كُتيّب التعلم العميق

ترجمة: سري السباعي

[مسوّدة غير مراجعة]

بسم الله الرحمن الرحيم

مُقَدِّمَة المُتَرجِمِ

[1]الحمد لله رب العالمين، والصلاة والسلام على محمد عبده ورسوله أما بعد:

فهذه ترجمة لكُتَيِّبٍ في التعلم المُحَوْسَبِ العميق المُرَكَّب لمؤلِّفُهُ "فرَنْشِسكُ فلُريت" وعنوانه بالكلزية:

The Little Book of Deep Learning François Fleuret

أحببت ترجمته مُقَدِّمًا به سلسلة أسأل الله تيسير خروجها في أصول علوم البيانات، والتعلُّم المُحَوسب، ومعالجة اللغات الطبيعية لسهولة الكتاب واحتوائه على الأصول المهمات والعناوين الرئيسيَّات في المجال نسأل الله الإعانة والتيسير.

وأدناه بعض المراجع الأساسية المُستعملة في ترجمة المصطلحات

- المورد الحديث
- مصطلحات المعلوماتية والرياضياتية من مجمع دمشق
- معجم سدايا
- معجم طعيمة

وأما ما اجتهدت فيه فعادة أنبّه عليه في الحاشية.

---

[1] شوم ع اب ج د و ح ط ي ك ل م ن س ع ف ص ق ش ت ث خ ذ ض غ ں ڡ ٯ ٮ ﺟ ھ ﺣ كلنسعصقتثخغجح يلنسصقشخضغںڡباجاحاطايا كامانا عافاصا قاشاتا خاضاظاغاںاڡابوھو نوطيولوم نوع طايا كاناسا عاصا قاشاتاثا خاظاغاںاڡابوج ھوو ح طيولوم نوسوع ڡوقو ر شوتوخ ذضوظوغ ب ج د و ح ط ي ل م ن س ف ص ق ش ت ث خ ذ ض ظ غ حـد حـد

فهرست الكتاب

مقدمة المؤلف

الفترة الحالية من تطور الذكاء المُصطنع[2] نُكِشتْ عندما برهن كَرِيزفسكي أن شبكة عصبية مُصطنعة مُصممة من عشرين سنة تفوق أعقد الطرق المُعتمدة[3] في معالجة الصور بفارق كبير فقط لما كانت ذات حجمٍ أكبر وتدرَّبت على مُعطيات أكثر.

هذا الفتح يشكُرُ وجوده لكروت الشاشة (وحدات معالجة الرسوميّات واختصارا الكروت) وحساباتها المتآزية عالية السرعة والدقّة المُطوّرة لتدريب الشعصبونات على الصور آنيًّا.

ومن وقتها وتحت جَناح "التعلم العميق" أباحت هياكل الشبكات وخُطَطُ، وعتاد تدريبها سامحةً نُمُوًّا أسيًّا في الحجم والكَمِّ المُعالج. ونتج عنه موجة تطبيقات ناجحة على مدار الفروع التقنية من مجال الرؤية الحاسوبية (صوريًّا)، ومجال الآليين (الروبوتات) إلى المعالجة الكلام الصوتي ثم من عام 2020 إلى تطوّر النماذج اللغوية العملاقة ذات القدرات التفسيرية العامّة.

ومع أن جوهر التعلم العميق سهلٌ فهمه إلا أنه يضمُّ علوما (مركبات أو أجزاء) أخرى من الجبر الخطيّ، والحُسبان، والإحتماليَّات، والاستمثال، ومعالجة الإشارات، والبرمجة والخرزمات، والحوسبة فائقة الأداء تجعلُ فهمه وتعلُّمه عسرًا.

وبدلا من الإحاطة والإطالة كان هذا الكتيب محدودا لفهم أهم أصول بعض أهم النماذج. مُثبتًا نجعَه وانتشاره بين ما يزيد على خمسمئة ألف قارئ في سنة واحدة، ورجاء إلَّم تحصل عليه من الموقع فافعل لنحصيَ عدد القراء https://fleuret.org/public/lbdl.pdf .

فرَنْشِسكُ فلُريت

19 مايو، 2024.

[2] هذا بدل المشهور من "الذكاء الاصطناعي"

[3] استعملت هذه الكلمة بدلا من قولهم: state of the art

القسم الأول: الأسس

الفصل الأول: التعلم المُحَوسب (التعلم الآلي)

التعلم العميق تاريخيا ينتمي إلى مجالٍ إحصائي كبير اسمه التعلم المُحَوسب جوهره مُهتم بطرق تعلم تمثيلات (أنماط سلوك) المُعطيات (البيانات)، هذه التقنيات أتت من الشبكات العصبية المُصطنعة وأما العُمق (جزء العميق من المصطلح) يبيّن أن النماذج هي سلسلة طويلة من التطبيقات المتراكبة تُعرف الآن بتحقيقها أداء أحسن.نمطية وجودة وتمدّد النماذج أطفر الأساليب الرياضية والبرمجية وأبرز التعلم العميق مجالا تقنيا فريدا واسعا.

التعلّم من البيانات

أسهل مثال لنموذج يُدَرب من مُعطيات معينة بشرط أن تكون لدينا س (أي نقطر على استعمالها)، هو صورة لوحة سيارة يُراد توقع ص منها وفي حالتنا توقع نص حروف اللوحة المكتوبة.

في كثير من الحالات الواقعية حيثُ س إشارة متعددة الأبعاد تُلتقط من بيئة غيرِ مُنحكمة تتعقد ويتعسر كتابة معادلة رقمية تقرنُ س بـ ص.

ولكن في قدرتنا أن نجمع مجموعة تدريب كبيرة م فيها أزواج (س$_{ن}$،ص$_{ن}$) وكذا أن نبنيَ نموذجًا وسيطيًّا ه. ما سبق اعتبره نصا برمجيًّا يربط متغيراتٍ مُتعلمة غ والتي تحاكي سلوك الدالّة وعليه فبِقَيِمٍ مناسبة غ* يصير لدينا مُتَوقِّعٌ جيّد. والجودة هنا تعني أنه لو كان لدينا س وأعطيَ لهذا النص البرمجي فإن القيمة $\widehat{ص}$=ه(س؛غ*) تحسبُ تقريبًا جيدا لـ ص المقرونة (المتعلقة) بـ س من بيانات التدريب الموجودة.

ترميز الجودة يصاغ عادة بدالَّة خطأ أو خسارة خ(غ) والتي تكون صغيرة عندما تكون ه(.؛غ) حسنةً (صالحة) على م. وعليه فإن تدريب (تعليم) النموذج يشمل حساب قيم غ* تستنقص (تستدني) خ(غ*).

غالب محتوى هذا الكتاب هو حول تعريف ه والتي في الأحوال الحقيقية هي تركيب من أجزاء معرفة.

المتغيرات المُتعلمة (المندربة) التي تركب غ يُطلق عليها اسمُ أوزان مُقتبساً اصطلاحها من متشابك أوزان الشبكة العصبية الحيوية. بالإضافة لهذه المتغيرات غالبا ما تقوم النماذج على متغيرات مُنحكمة[4] والتي تُحدد حسبَ المجال

[4] اختراعنا لترجمة hyper parameters ومنحكمة لأن الباحث هو الذي يتحكّمُ فيها وأما المندربة فهي نتيجة تدريب النوذج لا يدَ له فيها.

بخبرة سابقة أو أفضل التجارب السابقة أو بمحدودية المقدرة الحوسبيَّة وقد يُستمثلوا بطريقة أو بأخرى ولكن بأساليب مختلفة عن المستعمل في استمثال غ.

قاعدة (أساس) دالّة الانكفاء

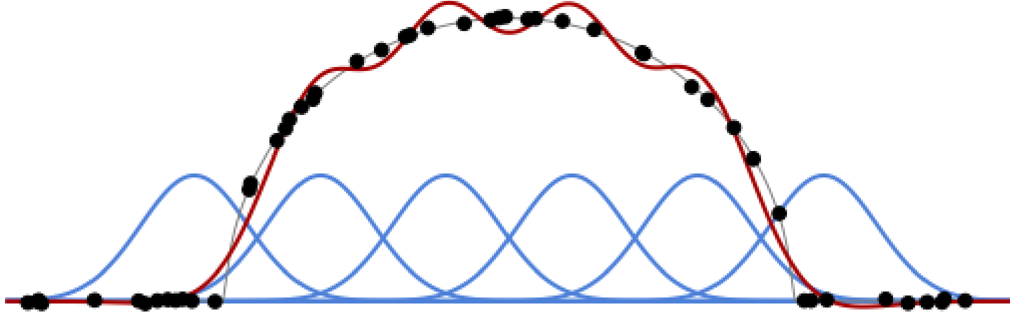
بمقدورنا تمثيل تدريب نموذجٍ ذي حالةٍ سهلة (بسيطة) حيثُ $\text{س}_{\text{ن}}$ و $\text{ص}_{\text{ن}}$ رقمان حقيقيَّان دالَّةُ خطئها هو متوسط مربع الخطأ

$$\text{خ}(\text{غ})=\frac{1}{\text{ن}}\ \text{مجـ}_{\text{ن}=1}^{\text{ن}}\left(\text{ص}_{\text{ن}}-\text{هـ}(\text{س}_{\text{ن}}؛\text{غ})\right)^{2}$$

حيث $\text{هـ}(\cdot؛\text{غ})$ تركيب خطيٌّ لقاعدة دوالّ معرّفة مسبقًا $\text{هـ}_{1}, \dots, \text{هـ}_{\text{ك}}$ مع $(\text{غ}_{1}, \dots, \text{غ}_{\text{ك}})$:

$$\text{هـ}(\text{س}؛\text{غ})=\text{مجـ}_{\text{ك}=1}^{\text{ك}}\ \text{غ}_{\text{ك}}\ \text{هـ}_{\text{ك}}(\text{س})$$

حيثُ ه(س$_{ن}$؛غ) خطيَّةٌ بالنسبة لـ الغينات $_{ك}$ و خ(غ) تربيعيّة بالنسبة للدالة ه(س$_{ن}$؛غ)، مع الخطأ خ(غ) تربيعي بالنسبة للغينات غ$_{ك}$ ونجد: غ* التي تستدني وتحل النظام الخطيّ. وانظر الرسمة التالية كمثالٍ مع أنوية جَوسيَّة كدالَّة ه$_{ك}$

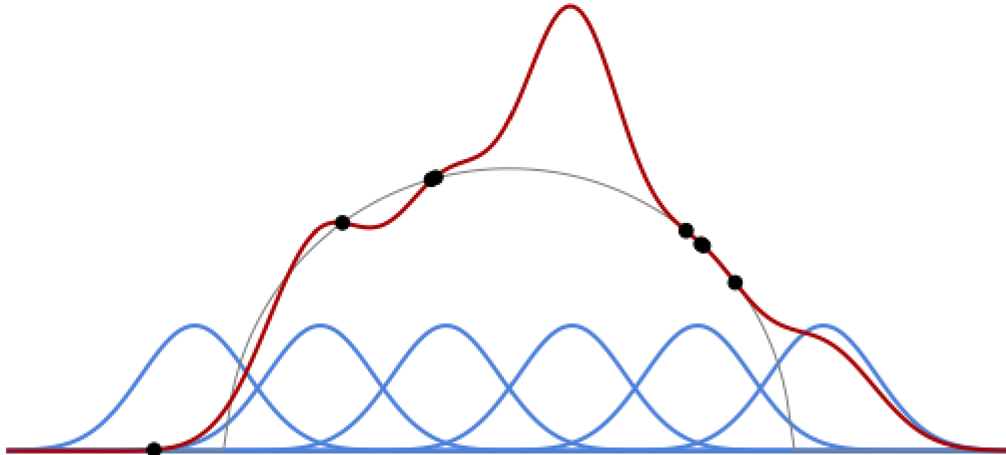


هذه الرسمة: مُعطًى قاعدة الدوالّ (المنحنيات الزرقاء)، وبيانات تدريب (النقاط السود)، يُمكننا حساب تركيب خطيٍّ أمثل يُمَثِّل (المنحنى الأحمر) لتقريب الأخير لمتوسّط مربع الخطأ.

التشبَّع، والتجوُّع[5]

هذه مسألة مهمّة وهي الموازنة بينَ سِعَة النموذج المرنة وقُدرتها على استيعاب المُعطيات المتنوّعة مع كميّة وجَوْدَةُ بيَانات التدريب (المُعطيات). إن كانت السعة غير فعالة (كافية) فالنموذج غير قادرٍ على استيعاب المُعطيات فيُنتج خطأً كبيرا أثناء التدريب وهذه هي مسألة التجوّع.

وعلى النقيض فعند كونِ المُعطيات غير كافية كما هو ممثل في الشكل التالي فإن النموذج يتعلم مميِّزات (صفات) مُحددة لعدد من أمثلة التدريب، منتجًا أداءً ممتازًا أثناء التدريب تلائم بُنية المعطيات العامّة ولكنها سيئة الأداء مع المعطيات الجديدة وهذه ظاهرة التشبع.



[5] هذه الترجمة للمصطلحين over fitting & under fitting

كمية بيانات التدريب (النقاط السود) صغيرة مقارنة مع سعة النموذج فالأداء الرقمي(التجريبي) للنموذج الملائم أثناء التدريب (المنحنى الأحمر) يعكس ملاءمةً سيئة للمنحنى الأصلي (بُنية البيانات الأصلية) (المنحنى الأسود النحيف)، وهذا متزامنٌ مع فائدته للتوقع.

اعلم أن جزءًا كبيرا من فن التطبيق لمجال التعلم المحُوسب هو في تصميم نماذج ليست مرنةً جدا ومع ذلك تلائم مُعطياتها. يكون هذا بتصنيع تحيزٍ استقرائي صحيح للنموذج تشير لبُنية مفيدة.

ومع أن الأساليب التقليدية مفيدة ومناسبة للنماذج الصغيرة فإن الأمور تتشوَّشُ مع الضخام ذات المتغيرات الكثيرة وسعةٍ ضخمة ومع ذلك تحسن التوقع وسنعود لذلك في فصلين قادمين [ان شاء الله].

فئات النماذج (أنواعها)

يُمكن ترتيب نماذج التعلم المحَوسب لثلاثٍ أساسيات:

- توقع (انكفاء): هو توقّع رقمٍ حقيقي (وقد يكون متجها) ص∋ح^ك، ومثالا الموقع الهندسي لجسمٍ لما تُعطى إشارةَ مُدخل س. وهذا تعميم عديد الأبعاد لما سبق في الفصل الماضي. بيانات التدريب تتكون من أزواجِ قيم المُدخلات وقيم المُخرجات الصحيحة.
- تصنيف: هدفهُ توقع قيمة من مجموعة محددة {1، ... ،ف} مثالًا الوسم (المُخرج) ص لصورة معينة س. كما في التوقع فالمُعطيات أزواج مُدخلات ومخرجات صحيحة وسمُه من مجموعته. الطريقة المعيارية لحل هذه المشكلة هي في توقع قيمة معيّنة لأحد المُخرجات، بحيثُ تكون هذه القيمة هي أكبر قيمة مُتوقّعة.
- نمذجة الكثافة: هدفها نمذجة إحتمالية كثافة دالة لمُعطيات م~س~. كمثال الصور. في هذه الحالة مجموعة المُعطيات مركبة من س~ن~ بدون مُخرج محدد يُتوقّع والنموذج المُدرَّب ينبغي أن يسمح لتقييم إحتمالية كثافة الدالّة، أو استخراج عيّنات من التوزيع أو كلاهما.

كلا التوقع والتصنيف عادةً يُطلق عليهما تعلم مُشْرف عليه إذ شرط هذا التدريب توفير المُخرج الصحيح المُتعلَّم منه من خبراء بشر متخصصين. وعلى الجانب الآخر فنمذجة الكثافة تعلم غير مُشرَفٍ عليه إذ إنها فعالة في التعامل مع المُعطيات بدون انتاجِ علاقة (اقتران) مع المُخرج الصحيح.

هذه الفئات (الأقسام) الثلاثة غير منفصلة (هي متداخلة) فإن التصنيف قد يُنظر لها كمسألة توقع أو سلسلة مقطعة لنمذجة كثافة كتصنيف متتابع. وهذه الأقسام لا تغطي كل الأحوال فالبعض يريد توقع كميات مركبة أو فئات عديدة، أو كثافة نموذج اشتراطي لمُدخل معيّن.

الفصل الثاني: الحساب العالي (المُجَوَّد)

من مبدأ التصميم (البرمجي) فإن التعلم العميق مداره على حوسبة أحجامٍ مهولة من البيانات. فَكَرتُ الشاشة (وحدة المعالجة الرسوميّة) عزف معزوفة نجاح المجال إذ سمح بهذا الحساب بعتادٍ منْشَرٍ.
أهميَّتهم ونتائجهم التقنية المشروطة بحسابها بشكلٍ فَعَّالٍ أجبرت المجال على حصرِ أصوات الرياضيات ومُبَرمجُ الطرق المميَّزة.

كروت الشاشة، ووحدة معالجة المرصوصات[6]، والدفعات

وحدات معالجة المرسومات أوَّلُ إنشائها كان لمعالجةٍ آنيّة لتوليف الصور، وذا يتطلب بُنًى متآزية شديدة الجودة والفعالية صادف موافقتها للنماذج العميقة. ومع زيادة استعمالهم في الذكاء المُصطنع فإن كروت الشاشة عُزِّزَت بأنويات مرصوصة ورقائق مُخصصة للتعلم العميق مثل وحدات معالجة المرصوصات من جوجل.
حوى كرت الشاشة بضع آلافِ وحدةٍ متآزية مع ذاكرتها السريعة. ولكن الحدَّ المانع عادةً ليس عدد الوحدات الحسابية ولكنَّها مسألة قراءة وكتابة العمليَّات في الذاكرة. أقصر رابط هو الذي بين ذاكرة المعالج وذاكرة الكرت وبالتالي فالمرءُ ينبغي ألا يُكثِرَ نسخَ البيانات بينَ الأجهزة. وعليه فبُنيَة (هيكليَّة) كرت الشاشة نفسها تتضمن مراحل (مستويات) من خابية الذاكرة، وهم أصغر وأسرع والحساب ينبغي أن يترَتَّب لتقلَّ النسخ بين الخوابي.
يُدركُ هذا تحديدا بترتيب المحسوبات في دفعاتِ عَيِّنات تُحشَر (تلائم) في ذاكرة الكرت كله وتُحسب بالتآزي. لما يدمج الباحث (المعامِل) العينة ومتغيرات النموذج كلاهما يُنقلان لخابية الذاكرة قرب وحدات الحساب الحقيقيَّة. استعمال نظام الدفعات يجعل نسخ أوزان (متغيرات) النموذج يحصل مرة واحدة بدلا من تكراره لكل عينة. تجريبيًّا الكرت يعالج دُفعةً محشورة (متلائمة) في الذاكرة بنفس سعرة حساب العيِّنة.
الكرت المعياري نظريًّا قمة أدائه تبلغ 10^{13}-10^{14} عملية نقطةٍ عائمة (عطائم) (FLOPs) في الثانية وذاكرتها تتراوح بين 8 إلى 80 جيجابايت عن 32 بِتًّا ولكن النتائج التجريبية تُظهر أن استعمال تشفير الـ16 بِتًّا أو أقل لبعض العمليات لا يُضعف الأداء. وسنعود للبُنى الضخمة في فصلٍ قادم.

---

[6] هي ما يُسمى بـTPU.

المرصوصات (المرصوفات)

كروت الشاشة وأُطر التعلم العميق كتَيرشٍ وجَكسٍ pytorch & JAX تتلاعب (تتعامل مع) الكمّيّات المعالجة بترتيبهم في مرصوصات وهي سلسلة أعداد رُتبت حول محاور متقطعة. وهم عناصر $ح^{ن1 \times ... \times ن د}$ والتي تعمم رمزَ المتجه والمصفوفة.

المرصوصات تُستعمل لتمثيل المُدخلات (الإشارات) المراد معالجتها والمتغيرات المُندرِبة للنموذج والكميات الوسيطة المراد حسابها. الأخيرة تُسمَّى مُفَعِّلات[7] إشارةً لتفعيل العصبونات.

مثالٌ السلاسل الزمنيّة عادة تُمَثَّل كمرصوصة أبعادها ت×د ولأسباب تاريخية تصير د×ت حيث ت هو المدة وال د هو بُعد المُدخلات عند كل خطوة زمنيّة، عادةً يشار لها بعدد القنوات. وكذلك الإشارةُ ذات بِنيةٍ ثنائية الأبعاد يُمكن تمثيلها كـمرصوصة أبعادها د×ط×ع حيث ط و ع هما الطول والعرض. الصورة ثلاثية الألوان (أحمر وأخضر، وأزرق ← حضرية) تجعل د=3، ويُمكن أن تمو أعداد القنوات إلى الآلاف في النماذج الكبيرة.

إضافة أبعادٍ أخرى تسمح بتمثيل سلاسل الأشياء. مثل خمسين صورة حضرية بُعدُ كلٍّ منها 32×24 تُضمُّ في مرصوصة بُعدها: 50×3×24×32.

مكاتب التعلم العميق تقدم عددا كبيرا من العمليّات التي تشمل أسس جبر الفضاءات[8] والتشكيل والاستخلاص وعمليات التعلم العميق المخصصة بَعضٌ منها يُذكر في الفصل الرابع [ان شاء الله]. تطبيق المرصوصات يفصل شكل التمثيل من تخطيط الذاكرة مع المعاملات في الذاكرة والتي تسمح بتشكيلات ومنقولاتٍ وسحب عمليات والتي تتم بدون نسخ المعامل وعليه بسرعة كبيرة.

تجريبيا نفترض أي محسوب يُمكن ضغطه في عملياتِ مرصوصةٍ أساسية والذي يتجنَّبُ أي حلقاتٍ غير متآزية على مستوى اللغة البرمجية وسوء إدارة الذاكرة.

بغض النظر عن قناعة الأدوات فإن المرصوصات أصل في منجزات الحوسبة الفعالة. كل الناس المنضمين في تطوير نموذج عميق عملي من مصممين ومهندسين ومكتباتها ونماذجها إلى هؤلاء العاملين على الحاسوب والكروت يدركون أن البيانات تُعامل معاملة المرصوصات. نتج عن ذلك شروط (محددات) محلِّيٍّ وكتل المنحلات سمحت كل العاملين في هذه السلسة بأن يأتوا بأحسن تصميم (أمثل).

---

[7] دوال التفعيل كناية عن تفعيل عصبونات المخ ومن صفاتها أيضا التحفيف أو الضغط. [مسودة سريعة تراجع] وترجمة logistic سوقية من معجم الرياضيات لبورفاين

[8] المشهور والمعروف الجبر الخطي وأنا أحب هذا لقوة دلالته على فحوى العلم

التدريب

كما وُصف وقُدم في الفصل الأول فالتدريب أصله استدناء (استصغار) دالة الخطأ خ(غ) والتي تعكس أداء النموذج المتَوَقِّع ه(٠؛غ) على مجموعة التدريب م.

وطالما أن النماذج معقدة شديدة التعقيد وأداءهم متعلق مباشرةً بكيف تُستصغر دالة الخطأ فذا أساس التحدِّي المتضمِّن تحدياتٍ رياضية وحوسبية (حسابية).

دوال الخطأ

من أمثلتها دالة حساب متوسط مربع الخطأ وهي دالة قياسية (معيارية أساسية) لتوقع القيم المستمرة.
أما عن نمذجة الكثافة، فالدالة القياسية فيها هي أرجحية البيانات. إن كانت ﻫ(س؛غ) مُستعملةً كمتناظمة احتمالٍ لغترميٍّ (لوغارتمي) أو كثافة لوغارتمية فإن دالة الخطأ (الخُطآن) هو معكوس مجموع قيمها على عيِّنات التدريب، والذي يمُثل أرجحية مجموعة المُعطيات.

الاعتلاج المتصالب[9]

للتصنيف فالخطة المُعتادة أن مُخرج النموذج متجه واحد ه(س؛غ)$_{\text{ص}}$ لكل صنف (فئة)، يُفسَّر على أنه لوغارتمُ الاحتمال غير النظيمي (المُنَظَّم) أو السجلات[10].

فإن المُدخل س ومقابله المُخرج ص (الصنف المُعيَّن) المراد توقعه فنحسب من الدالة ه تقريبا الإحتمالات البعدية:

$$\hat{P}(Y = y \mid X = x) = \frac{\exp f(x;w)_y}{\sum_z \exp f(x;w)_z}.$$

$$\hat{\text{إ}}(\text{س} = \text{س،ص} = \text{ص}) = \frac{\text{هـ}^{\text{د(س؛غ)}_{\text{ص}}}}{\text{مجـ}_{\text{ز}}\,\text{هـ}^{\text{د(س؛غ)}_{\text{ز}}}}$$ [11]

---

[9] هذه الترجمة اجتهاد مني لكلمة "cross entropy" فالأولى وجدتها استُعملت في المعجم بالتصالب وفي معجم طعيمة متقاطعة والثانية وجدتها مُعربة "إنتروبيا" وهي كذلك في معجم طعيمة ووجدت في موقع المعاني ترجمتها بالاعتلاج فأحببت استعمال الكلمة العربية

[10] أول من وجدت ترجم كلمة logits الأستاذ طعيمة في كتابه معجم الذكاء الاصطناعي بالسجلات صـ131 "متجه التنبؤات الأولية (غير المطبعة non-normalized) التي يولدها نموذج التصنيف ..." وأما عندي فالأقرب ترجمتها بشيء من مرادفات الخام إما خوامّ أو مُحُض أو نيئات أو نُقاة.

[11] /فوق {إ}{/و.}(س=س،ص=ص)=/على{/خط.فو {"ه"}{/خط.تح{د(س؛غ)}{ص{يسار}}{يسار}}{/مج{""}{ز}/خط.فو {"ه"}{/خط.تح{د(س؛غ)}{ز{يسار}}{يسار}}

هذا التعبير (المعادلة) عادةً يُسمى مُحَصِّيًا[12] وبصياغة أدق: الاستقصاء الناعم للسجلات.

لتكون متسقًا مع التعليل (التفسير) النموذج ينبغي أن يُدَرَّب لاستقصاء احتمالات الفئات الحقيقية وعليه استقصاء (استعظام) الاعتلاج المتصالب:

$$\text{خ}(\text{غ}) = -\frac{1}{\text{ن}} \text{ مجـ } \hat{\text{إ}} \left(\textbf{ص}=\text{ص}_{\text{ن}} \mid \textbf{س}=\text{س}_{\text{ن}}\right)$$

$$= \frac{1}{\text{ن}} \sum_{\text{ن}}^{1} - \text{لوغا} \frac{\text{هـ}^{\text{د}(\text{س؛ غ})_{\text{ص}}}}{\text{مجـ}_{\text{ز}} \text{هـ}^{\text{د}(\text{س؛ غ})_{\text{ز}}}}$$

حيث الجزء المجموع هو دالة خطأ الاعتلاج المتصالب $\text{خ}_{\text{إم}}(\text{ه}(\text{س}_{\text{ن}}\text{؛غ}),\text{ص}_{\text{ن}})$

الخطأ التبايني[13]

في أحوال معينة وإن كان المتَوقع قيمة حقيقية (مستمرة) فالإشراف عليها يتطلب ترتيبا للثوابت. والمجال الأمثل للتمثيل عليه هنا هو التعلم القياسي حيثُ المقصود فيه وهدفه تعلم قياس المسافة بين العيِّنات مثالا $\text{س}_{\text{أ}}$ من فئة معنوية معيَّنة أقرب إلى أي عيِّنةٍ $\text{س}_{\text{ب}}$ من نفس الفئة مقارنة بعينة $\text{س}_{\text{ت}}$ من فئة أخرى. خذ مثالا $\text{س}_{\text{أ}}$ و $\text{س}_{\text{ب}}$ يُمكن أن تكونا صورةً لنفس الشخص وتكون $\text{س}_{\text{ت}}$ صورة لشخص آخر.

[12] هذه ابتكرتها من طبيعة عمل الدالة وهي تحويل "السجلات" أو المحُض إلى توزيع احتمالي بشروطه.

[13] هذا المصطلح لم أجد من ترجمه "Contrastive loss" والكلمة الأولى يُمكن أن تكون تشابهيًّا

الطريقة المعيارية (القياسية) لهذه الأحوال هي في استدناء (استصغار) دالة الخطأ التبايني، حيثُ فيها كمثالٍ أن مجموع الثلاثي ($س_{أ}$، $س_{ب}$، $س_{ت}$) بحيثُ أن تتساوى $ص_{أ}$ = $ص_{ب}$ ≠ $ص_{ت}$ ل

أقصى (0, 1 - ه($س_{أ}$، $س_{ت}$؛ غ) + ه($س_{أ}$، $س_{ب}$؛ غ))

فهذه الكمية ستكون موجبةً إلا إذا:

ه($س_{أ}$، $س_{ت}$؛ غ) ≤ 1+ ه($س_{أ}$، $س_{ب}$؛ غ).

هندسة دالة الخطأ

عادة فدالة الخطأ التي تُستدنَى (تُستصغَر) في التدريب ليست هي القيمة (الكمية) الحقيقية التي نريد استمثالها بشكل عام، ولكن ما يحصل هو ايجادُ أفضل متغيرات للنموذج بسهولة. مثالًا في التصنيف، فإن الاعتلاج المتصالب هو دالة خسارته مع أن المقياس الحقيقي لأداء النموذج هو معدل خطأ التصنيف، ولكن لانعدام انحداره معلوماتيًّا وهو أساس كما سيأتي في الفصل القادم [ان شاء الله].

وممكنٌ أيضا اضافة أجزاء لدالة الخطأ تعتمد على متغيرات مُندربة في النموذج لتمييز خصائص معينة.

تنظيم اضمحلال الوزن مثل إضافة جزءٍ (قيمة) لدالة الخطأ باطرادٍ مع مجموع مربع المتغيرات. هذه تُفَسَّر على أن لدينا قبليٌّ جوسيٌّ بيزيٌّ للمتغيرات والذي يُفضل القيم الأقل بالتزامن مع تقليل أثر البيانات. السابق يُقلل (يُنقص) الأداء على بيانات التدريب ولكنه يُضاءل (يُقلل) الفراغ (الفجوة) بين التدريب والجُدُد غير المسبوقين (المشاهدين).

النماذج المتحادرة[14]

هناك فئة أساسية من الطرق خاصةً للتعامل مع التسلسلات (المتتاليات) المُتقطِّعة (المنفصلة) في معالجة اللغة والرؤية المَحُوسبة وهي نماذج التحادر (النماذج المتحادرة).

قاعدة سلسلة الإحتمالات

بعض النماذج تستعمل قاعدة السِّلسلة من نظرية الإحتمال:

ح[15](س$_1$=سـ$_1$,س$_2$=سـ$_2$, ... , س$_ر$=سـ$_ر$)=

ح(س$_1$=سـ$_1$) × ح(س$_1$=سـ$_1$ | س$_2$=سـ$_2$) ... × ح(س$_ر$=سـ$_ر$)| س$_1$=سـ$_1$, ... س$_{ر-1}$=سـ$_{ر-1}$).

ومع أن هذا التحليل (التفريق) مسموح به لأي متتالية عشوائية من أي نوع فإنه بالخصوص فعال عندما تكون الإشارة (يكون المُدخل) المتعامل معه سلسلة (متتاليًا) من الوحدات[16] من محدودات (مُفردات محدودة) }1, ... ك{.

ومع إضافة وحدة Ø تكون لغير المُعَرَّفات يُمكننا تمثيل الحالة (المُدخل) }س$_1$=سـ$_1$, ... س$_ر$=سـ$_ر$} متجهًا (س$_1$, ... س$_ت$, Ø , ... , Ø).

ثم النموذج

ه:}1,Ø, ... , ك{^{م} ← ح^ك

والذي بالمُدخل المعين يحسب متجها ل $_ت$ من مُحُض ك تشير إلى:

ح(س$_ر$| س$_1$=سـ$_1$, ... س$_{ر-1}$=سـ$_{ر-1}$)

والتي تسمح بأن تُختار الوحدة الجديدة بناءً (مُعطًى) على القديمات.

قاعدة السلسلة تضمن بأنَّ اعتيان ت وحدات (يعني وحدات عددها ت) س$_ت$ واحدةً في المرة الواحدة مُعطًى العينات السابقة (س$_1$, ... س$_{ت-1}$) فنحصل على متتالية تتبع التوزيع المُشترك وهذا هو نماذج التوليد المتحادرة.

تدريب نموذجٍ كهذا يكون باستنقاص (باستدناء/باستصغار) مجموع على متتاليات التدريب على الزمن لدالة خطأ الاعتلاج المتصالب

---

[14] هذه الترجمة لم أُسبق لها لمصطلح: "autoregressive" وسببه أنها تحادر مع مُخرجها وتعيده مدخلا. وأخيرًا رجحت "مجترة"

[15] ح هذه تعبر عن الاحتمالية

[16] هذه ترجمة لـ"token" ولعلي استعملت لفظا أعم من الكُليمات وهي في سياق النماذج اللغوية أو الوحدات اللغوية أو الأجزاء وهذه الكلمة تحتاج تحقيقا في استصلاحها فهي مُستعملة بشكل كبير ومتداخل. فمن ذلك رمز، أو قطعة، وأصل المعنى من المورد الأكبر علامة أو رمز أو خِصيصة وغيره. وأقترح أيضا مُتَرِّكات أو بُتوت (جمع بت) وبتاوتة وكذلك قطع وأقسام

$\text{خ}(\text{ه}(\text{س}_1, \dots \text{س}_{\text{ت}-1}, \emptyset, \dots, \emptyset), \text{س}_\text{ت})$

والذي يُمثل (يصوغ) [يعادل صوغًا أو صياغيا] لاستعظام (استكبار) أرجحية صواب كل $\text{الس}_\text{ت}$.

القيمة المُعتاد مراقبتها ليست الاعتلاج المتصالب نفسه بل الارتباك (حيرة)، والذي يُعرف على أنه أس الاعتلاج المتصالب. ويشير إلى عدد القيم من التوزيع المنتظم لها قيم الاعتلاج نفسها, وذا أسهل في التفسير.

النماذج السببية[17]

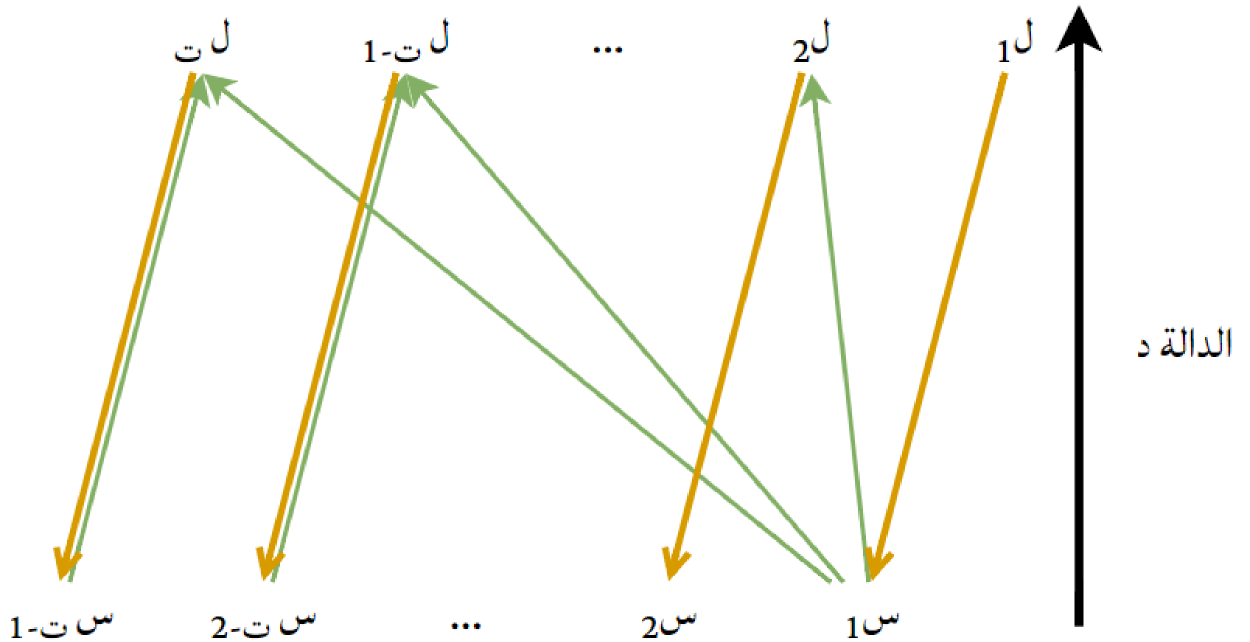
في عملية التدريب وصفنا ضرورة مُدخل مختلف لكل عند كل ت والكتلة الحوسبية المطلوبة لـ ت<ت* والتي تتكرر لـ ت*. هذا غير فعال أبدا حيثُ ت تكون بالمئات إلَّم يكن آلافًا.

الخطة المعيارية (القياسية) لحل هذه المشكلة هي في تصميم نموذج ه يتوقع كل مُخرجات المُحُض $\text{ل}_1, \dots \text{ل}_\text{ت}$ مرة واحدة وعليه:

$$\text{ه}: \{1, \emptyset, \dots, \text{ك}\}^{\text{م}} \rightarrow \text{ح}^{\text{ت}\times\text{ك}}$$

ولكن مع بُنية الحوسبة التي تحسب مُحُضَ لِ لـ $\text{س}_\text{ت}$ تعتمد فقد على ما قبلها $\text{س}_1, \dots \text{س}_{\text{ت}-1}$.

هذا النموذج يُسمى سببيا حيثُ أنه يشير في هذه الحالة حالة المتتاليات المؤقتة إلى عدم تأثير المستقبل على الماضي كما هو موضع في الصورة



النموذج المتحادر د يُسمى سببيا إن كان الخطوة الزمنية سِ لمتتالية المُدخل تقايس المحُض المتوقعة لِ فقد إن كان س<ت، كما هو مشار له بالسهم الأخضر. هذا يسمح بحساب التوزيعات على كل الخطوات الزمنية في سحبة واحدة في التدريب. في الاعتيان فال لِ وال سِ يُحسبان تتابعيا، فإن آخر عينة تُحسب من السابق كما هو موضح بالسهم الأصفر.

---

[17] ينبغي مراجعة هذه التسمية

المُقَسِّمات

تفصيل تقني مُهم عند التعامل مع اللغات الطبيعية هو تمثيلها بكُليمات (...) ويتم ذا بعدة طرق بدءًا بعدد محدود من الرموز المُفردة وانتهاءً بكلمات كاملة. عملية التحويل من وإلى التمثيل الكُليماتي (الوحدوي) يتم بحزمة مُنفصلة فاعلها اسمُه مقسم.

من الطرق الرئيسية (الأساسية) هو المزوج البِتّي والذي يبني كُليماته هرميا بدمج مجموعات المحارف محاولا الحصول على كُليمات تُمثل جُزءًا من الكلمات بأطوال مختلفة بتكرارات متشابها، موصلا الكُليمات للأجزاء الأكثر تكرارا بنفس الوقت مع النوادر.

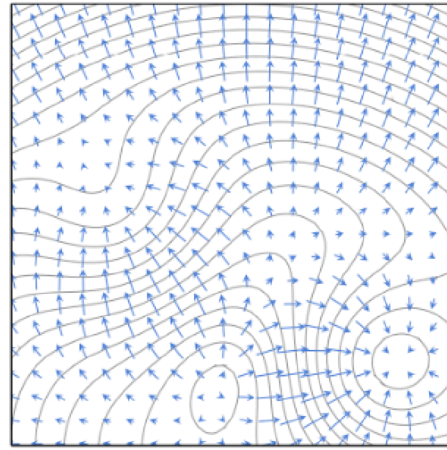
التدرج الاشتقاقي[18]

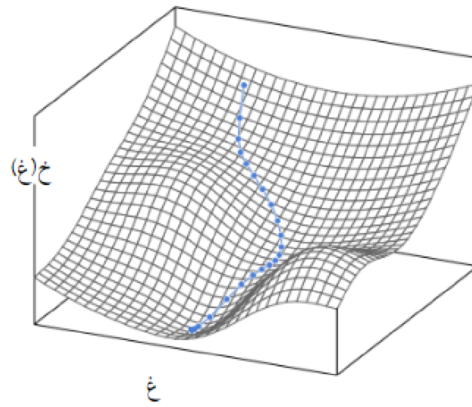
بغض النظر عن الحالات الخاصة كحالة الانحدار الخطي والتي يُمكننا فيها حساب المتغيرات المُثلى غ* فإن الغالب أن لا صيغة مباشرة لحلِّهم. ففي المُجمل العام فإن التقنية (الأداة أو الطريقة) المُستعملة في استدناء (استنقاص) [تقليل خطأ] الدالة هي التدرج الاشتقاقي. تبدؤ بتعبئة المتغيرات بقيم عشوائية $غ_0$ ومن ثَمَّ في كل لفة (كرة أو موجة) تُحسِّنهُ بخطواتٍ اشتقاقية تحوي كل منها مُشتق المخطوءة (دالة الخطأ أو الخسارة المخسورة) بالنسبة للمتغيرات وطرح جُزءٍ منها:

$$غ_{ن+1} = غ_{ن} - ذ \nabla خ_{غ}(غ_{ن})$$

[18] شوم ع اب ج د و ح ح ط ي ك ل م ن س ع ف ص ق ش ت ث خ ذ ض غ ب ف ق ب ج ه ح ك ل ن س ع ص ق ت ث خ غ ج ح ي ل ن س ص ق ش خ ض غ ب ق باجاحاطايا كامانا عافاصاقاشاتاخاضاظاغاباقا ب ه ه و ن ط ي ل و م ن ع طايا كانا ساعاصاقاشاتاثاخاظاغاباقابوج ه ه و ح ط ي ل و م ن س و ع ف و ق و ب ش و ت و خ ذ ض و ظ و غ ب ج د و ح ط ي ل م ن س ف ص ق ش ت ث خ ذ ض ظ غ حـ د حـ د

هذه المنهجية تشير إلى حركةٍ مُقدرة مسافةً في اتجاه المحلِّ الذي يقلل خ(غ) كما هو موضح في الشكل





في كل نقطة غ فتدرج ∇خ(غ) هو الاتجاه الذي يعاظم (يستعظم/يستكبر) خ معامدًا لمستوى المنحنيات (الأولى). التدرج المتناقص يقلل (يستصغر) خ(غ) تكراريا بطرح جزءٍ من التدرج في كل خطوة منتجا اسقاطا يتبع أعمق (أدنى) انحدار (الثانية)

معدل التعلم

المتغير المنْخَفِر ذ يُسمى بمعدل التعلم وهو قيمة موجبة تنظم سرعة انتهاء الاستنقاص (التقليل) وينبغي اختيارها بحذر.

لو كانت صغيرة جدا فالاستمثال يكون بطيئا على أحسن تقدير وقد يُحبس (يعلق) في دُنَيَوِيَّةٍ محلية باكرا. وعند العكس فالاستمثال يتردد حول محليٍّ جيد ولن يتقارب له كما سيأتي قد يعتمد على عدد التكرارات ن.

التدرج الاشتقاقي العشوائي

كل دوال الخطأ المُستعملة في حقيقي الممارسة تُمثل كتوسط خطأ مجموعة صغيرة من العينات أو واحدة:

$$خ(غ) = \frac{1}{ن} مجـ_{1}^{ن} خـ_{ن}(غ)$$

حيثُ $خـ_{ن}(غ)$ = ل(د(سن؛)،صن) لـ ل محدد وتدرجها يكون:

$$\nabla خ_{غ}(غ) = \frac{1}{ن} مجـ_{1}^{ن} خـ_{ن\nabla غ}(غ).$$

مُنتجا التدرج الاشتقاقي المساوي تماما للمجموع في المعادلة السابقة الثقيلة حوسبيا ومن ثَمَّ نُغيِّر (نحسن أو نطور) المتغيرات باستعمال المعادلة من القسم السابق.

الانتشار الخلفي[19]

استعمال التدرج الاشتقاقي يتطلب متوسِّطات تقنية لحساب ∇خ~غ~(غ) حيثُ
خ= خ(د(س؛غ),ص). مُعطى ه و خ كلاهما تركيب من عمليات مرصوصات أساسية وكأي مصوغ رياضي
قاعدة السلسلة من الحُسبان التفاضلي تسمح (تبيح) الإتيان بتعبير له.
طلباً لتسهيل الترميزات فلن نحدد في أي نقطة يُحسب المتدرِّج إذ السياق يُوَضِّحه.

[19] هذه الكلمة ترجمتها مختلفة في بعض المصادر فمنهم من قال انتشار عكسي ومنهم من قال خلفي

التمرير الأمامي والخلفي

افرض حالةً يسيرة من التطبيقات المُركَّبة:

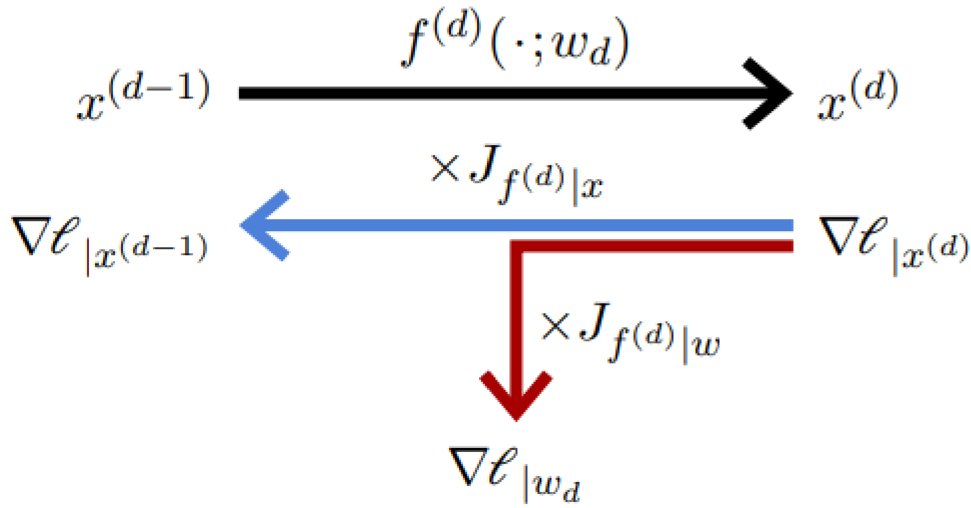
$$\text{د} = \text{د}^{(\text{ب})} \circ \text{د}^{(\text{ب-1})} \circ \ldots \circ \text{د}^{(1)}$$

المُخرَج من ﻫ(س؛غ) يُمكن حسابه بالبدء من س $^{(0)}$ =س ونطبِّق تكراريا:

$$\text{س}^{(\text{ب})} = \text{ﻫ}^{(\text{ب})}(\text{س}^{(\text{ب-1})}؛\text{غ}_{\text{ب}})$$

بأن تكون س $^{(\text{ت})}$ آخر قيمة.

قيمة المُحدد (السُّلَّمي أو العددي) المفردة للقيم الوسيطة س $^{(\text{ت})}$ تسمى عادة مُفَعِّلات إشارة لمُفعلات العصبونات، والقيمة ت هي عُمقُ النموذج والتطبيق المُفرد د $^{\text{ت}}$ يشير إلى الطبقات كما سيأتي وأما حسابهم التتابعي يُسمى التمرير الأمامية.



(أعتذر عن عدم ترجمة المعادلات هنا مؤقتا) مُعطى النموذج ﻫ = د $^{(\text{ب})} \circ$ د $^{(\text{ب-1})} \circ \ldots \circ$ د $^{(1)}$ فالانتشار الأمامي يَحسب المُخرجات x^d للدالة f^d بالترتيب (الأعلى الأسود). وأما التمرير الخلفي فيحسب المُتدرِّجات لدالة الخطأ بالنسبة للمُفعلات x^d (الأسفل الأزرق) وأما المُتغيرات w_d (الأسفل الأحمر) فتمريرهم الخلفي يكون بضربهم باليعقوبيَّات.

وعلى العكس فالتدَّرج ∇خ$_{|\text{س}^{(\text{ت-1})}}$ لدالة الخسارة بالنسبة للمُخرج س $^{(\text{ت-1})}$ لـ ﻫ $^{(\text{ت-1})}$ هو جداء المشتق ∇خ$_{|\text{س}_{\text{ت}}}$ بالنسبة للمُخرج ﻫ $^{(\text{ت})}$ مضروبا باليعقوبية ع $_{\text{د}(\text{ت-1})|\text{س}}$ لـ ﻫ $^{(\text{ت-1})}$ بالنسبة لمتغيِّرها س. وعليه فالمُتشقات بالنسبة لمُخرجهم لـ ﻫ $^{(\text{ت})}$ يُمكن حسابهم بارتداد خلفي بدءًا من ∇خ$_{|\text{س}(\text{ت})}$ = ∇خ$_{|\text{س}}$ والمشتق المهتمِّ به للتدريب الذي ∇خ$_{|\text{غ}(\text{د})}$ يُضرب بيعقوبيِّ ع $_{\text{د}(\text{د})|\text{غ}}$ بالنسبة للمتغيرات.

هذا الحساب المتكرر للمشتقات بالنسبة لمُفَعِّلاتهم الوسيطة مجموعا مع للمشتقات بالنسبة لمتغيرات الطبقات يُسمى بالمرور الخلفي (انظر الصورة السابقة) وارتباط هذا مع منهجية الانحدار الخطي يُسمى بالانتشار الخلفي.

تطبيقيا (تجريبيا) دقائق ترميز (كتابة النص البرمجي) هذان الانتشاران الخلفي والأمامي تُخفى عن المبرمجين. فأُطُرُ ومكتبات التعلم العميق طُوِّرت لتحسب آليا سلسلة عمليات الاشتقاق.

أحد الحزمات المفيدة تُسمى Autograd والتي تتبع مرصوصة العمليات وتبني لحظتها (آنها) مركبات عمليات المشتق ونشكر ذلك.

استعمال المصادر (الموارد)

بالنسبة للكلفة الحوسبية كما سنرى فالمحسوبات تصير عمليات خطية كل منها يطلب جداءًا مصفوفيا للانتشارين وجُداءين لليعقوبية في الخلفي جاعلًا الأخير بضعف الكِلفة.

الذاكرة المطلوبة في استعمال (استعمال النموذج بعد تدريبه) تساوي تقريبا لأشد الطبقات المُفردة. أما للتدريب فالانتشار الخلفي يتطلب إبقاء المُفعلات محسوبةً ضمن الانتشار الأمامي لنحسب اليعقوبيات والذي ينتج في استعمال نامٍ يطرد مع عُمقِ النموذج. توجد تقنيات لتهذِّبَ استعمال الذاكرة إما بالاعتماد على الطبقات المُنعكسة أو باستعمال نقاط حفظ والتي تحفظ المُفعلات لبعض الطبقات وتعيد حساب البواقي وقتها مع انتشار أمامي جزئي خلال الانتشار الخلفي.

اضمحلال المشتق

مشكلة تاريخية جوهرية عند تدريب الشبكات الكبيرة (الضخمة)عند انتشار المشتق خلفيا عبرَ معامل فقد تضاعف بعامل ضربي والذي قد ينمو أو يتضاءل أسيا لما يمر بعدة طبقات. طريقة معيارية أساسية لاجتناب ذلك (هذا الانفجار) هي قص نظيم المشتق والذي يحتوي على إعادة معامل المشتق ليُثَبِّت نظيم على عتبة معينة.

عند تناقص (تضاءل) المشتق أسيا فذا يُسمى اضمحلال المشتق (اختفاء) وقد يحيل التدريب مستحيلا أو في حالاته الوسيطة أن يجعل بعض طبقات النموذج تتغير بسرعٍ مختلفة مقللا تكيّفها.

وكما سنرى في الفصل القادم فإن عدة طُرُق (أساليب) طُورت لتجنب حدوث ذلك, عاكسةً تغيرا في منظور مهم في نجاح التعلم العميق: فبدلا من محاولة أساليب الاستمثال نوعيًّا فالمجهود انصب على هندسة النماذج نفسها لتكن مُنمَثلة (قابلة للاستمثال).

تكلفة (قيمة) العمق

كما مُصطلح "التعلم العميق" يقتضي أن النماذج المفيدة مركبة من تركيب تطبيقات طويل. تُدَرَّبون بالانحدار المتدرج منتجا علاقة معقدة من التطبيقات ولو كانَ متراكبا ومحليا.

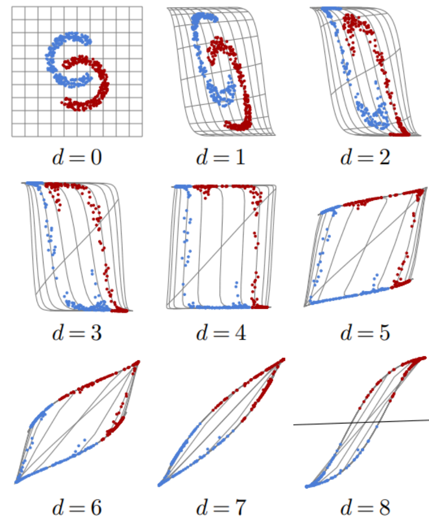
يُمكننا تمثيل (ترسيم) ذا السلوك بنموذج بسيط ح^2←ح^2 والذي يحتوي ثمان طبقات، كل منها تجدي مُدخلها مصفوفة 2×2 وثم تطبق دالة جتاه Tanh على كل مركب، ثم آخر طبقة للتصنيف. هذه النسخة المبسطة من الشعصبونات (شبكة العصبية) كما سنرى.

إن دربنا نموذج بالانحدار المتدرج (المتدارج) والاعتلاج المتقاطع على دُمية تصنيف ثنائي (الصورة التالي الأعلى يسارًا) فالمصفوفات شاركت في تكييف في تشويه الفضاء حتى ينجح التصنيف، والذي يقتضي أن البيانات صنعت تفريقا خطيا قبل آخر عملية تآلفية (الصورة التالي أسفلها يمينا).

والمثال يُعطي لذعةً لكيفَ تعمل النماذج العميقة. ولكنها مضللة أحيانا بسبب قلة الأبعاد سواءً في المُدخل أو عملية التمثيل الداخلي. وكل شيء هنا في بعدين تسهيلا على التمثيل حيثُ في النماذج الحقيقية في عوالي الأبعاد فالحرية أوسع.

الدليل العملي (الرقمي التطبيقي) مجمعا على عشرين سنة يثبت أن النماذج ذات الأداء المعياري (العالي المميز) في كل المجالات في النماذج التي فيها عشرات الطبقات مثل الشبكات الجارية (المتفرعة) أو المحولات.

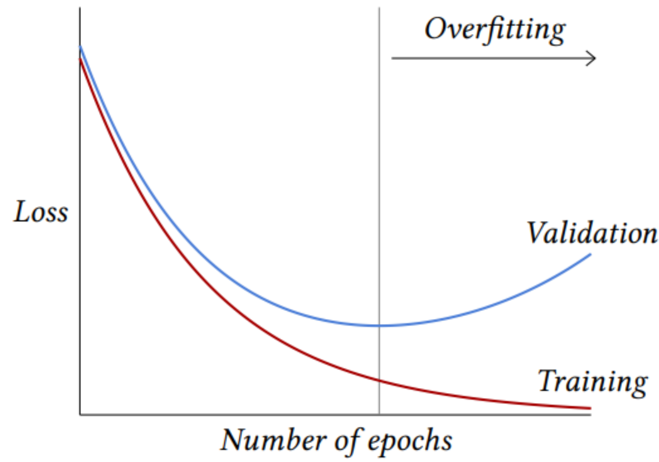
النتائج النظرية تُظهر أن لحوسبة محدودة الميزانية أو عدد المتغيرات وزيادة العمق تؤدي إلى تعقيد لنتائج التطبيق



كل رسمة تظهر تشوه (تراجع) الفضاء والنتائج الموضعية للتطريف في بعدين بعد عدد طبقات (هنا 8) بدءًا بالمدخل نفسه (أعلى اليسار) وانتهاء بالخط الفاصل في آخر رسمة (أسفل اليميني) تظهر آخر تآلف.

أصول التدريب (أساليبه)

تدريب شبكة عصبية يتطلب تعريف "أصل أو أسلوب" ليستقصي الحسبنة والبيانات بأعلى أداء على الجديد.
وكما مرَّ معنا الأداء على عينات التدريب قد يُضلل وعليه ففي أيسر حالٍ لا بُدَّ للمرء من عينتين واحدة للتدريب والأخرى للإختبار.
ومعهُ فهناك عادةً متغيرات مرنة (معاملات ضبط) للتلاعب بها (تكييفها) المرتبطات بهيكلية النموذج: معدل التعلم وحدود التنظيم (أي مُنَظِّمات التعليم أو ضابطاتها) في الخِطأة (دالة الخطأ) ولذلك فيُحتاج لمجموعة تحقق تنتج من عيِّنةٍ متقاطعة مسحوبة من مجموعة التدريب والاختبار لدعم أحسن تركيبة (خلطة).



بينما يتقدم التدريب فإن أداء النموذج يُراقب بالأخطاء، فخطأ التدريب ينبثق من تحسين العملية ويتنازل بينما خطأ التحقق يُقاس على عينة مختلفة ليدعم (يُسعف أو يؤازر) التدريب لكيلا يتشبع. فعندما يحسب هيا كل عشوائية في التدريب (يتشبع) فخطأ التحقق يتزايد.

وتناقضيا فقد تعاني (أي النماذج) من تشبع جسيم بسبب حجمها والنماذج الكبيرة (الضخمة) يحسن تدريبها كلما تقدم التدريب. وهذا يرجع لمبدأ التحيز الاستقرائي[20] للنموذج والذي يُعتبر الدافع الأساسي للتحسين عند وجود مجموعة تدريب مثالية[21].
أحد الخيارات التهيئية (التصميمية) هو جدول معدل التعلم أثناء التدريب وهي خصيصة (تغير) معدل التعلم أثناء التدريب كُلَّ لفة للمشتق المتدرج. الأسلوب العام (الأصل) هو أن يكون المعدل كبيرًا لكيلا ينسجن (يُحتجز أو

[20] يحتاج مراجعة لعدم تكلم المعاجم العربية عنه (وانظر ما كتبته في ورقتك لأنك ذكرته في باب سبب تعلم النماذج)
[21] Reconciling modern machine learning practice and the bias-variance trade-of

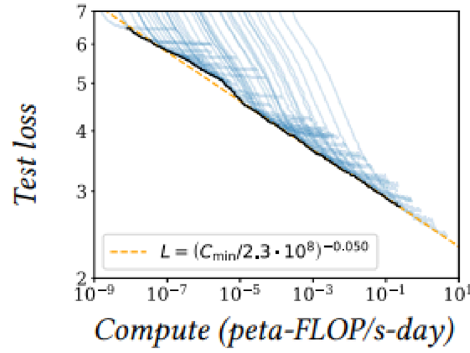
يعلق) في محلية صُغرى مبكرا، ثم تتقلص (تتصاغر) بحيثُ لا يتقافز (يرتدد) حول محلية جيدة في حيّز ضيق في فُرجة المخِطأة (دالة الخطأ).
تدريب نماذج ضخمة جدا يتطلب (يستهلك) شهورا من مجهود الكروت الشديدة وله تكاليف غالية تبلغ أبلافا (ملايين) وعلى حجم كهذا فالتدريب يتضمن تفاصيل تصب كلها في تحركات الخطأ.

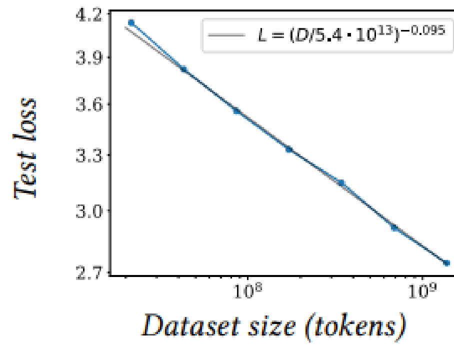
المعيرة

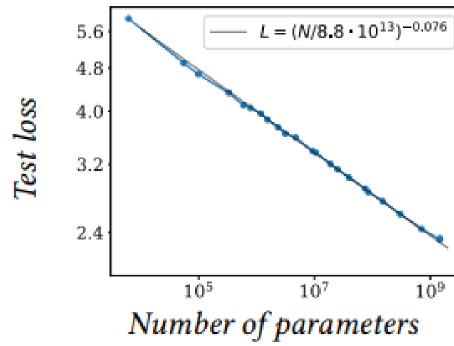
من المفيد استعمال نموذج مدرب لما يُسمى مهمة سُفلى (جزئية). لأسباب منها حجم البيانات الأصلية مهول بينما أداءهم على المهمات الجزئية محدود فالمهمتان التي تتشاركان هيكلًا احصائيا متشابها كفاية فالأول يُعطي الثاني تحيزا استقرائيا جيدا. وقد يكون محدودا لكلفة التدريب والتي تُحل باستخدام أنماط مُرمَّزة في نموذج جاهز (موجود).
استعمال نموذج مُدَرَّب لمهمة محددة يتم بالمعيرة والتي هي تدريب عادي على مهمة جزئية ولكنها تبدأ من أوزان مُدربة بدلا من انطلاق عشوائي (في التدريب العاديّ).
هذه خُطة (منهجية أو أسلوب أو مبدأ) غالب تطبيقات الرؤية الحاسوبية (المُحَوسبة) والتي عادة تستعمل نموذجا مدربا على تصنيف بيانات ألكسنت وأيضا فهو كيف النماذج اللغوية المُدربة تُحول لمدردشة (أشباه مساعدات) لإنتاج (توليد) محادثات تفاعلية.
سنعود لتقنيات (أساليب) التعامل مع محدودية المصادر في الاستدلال (التوليد) والمعيرة في الفصل الثامن.

فوائد التضخيم (التحجيم)

هناك تراكم في النتائج التجريبية (الرقمية) تُظهر أن الأداء مُقاسًا بالمخطأة على المُخْتَبِرات يتحسن مع كمية البيانات طردًا حسبَ قوانين التحجيم تزامنا مع تزايد حجم النموذج.

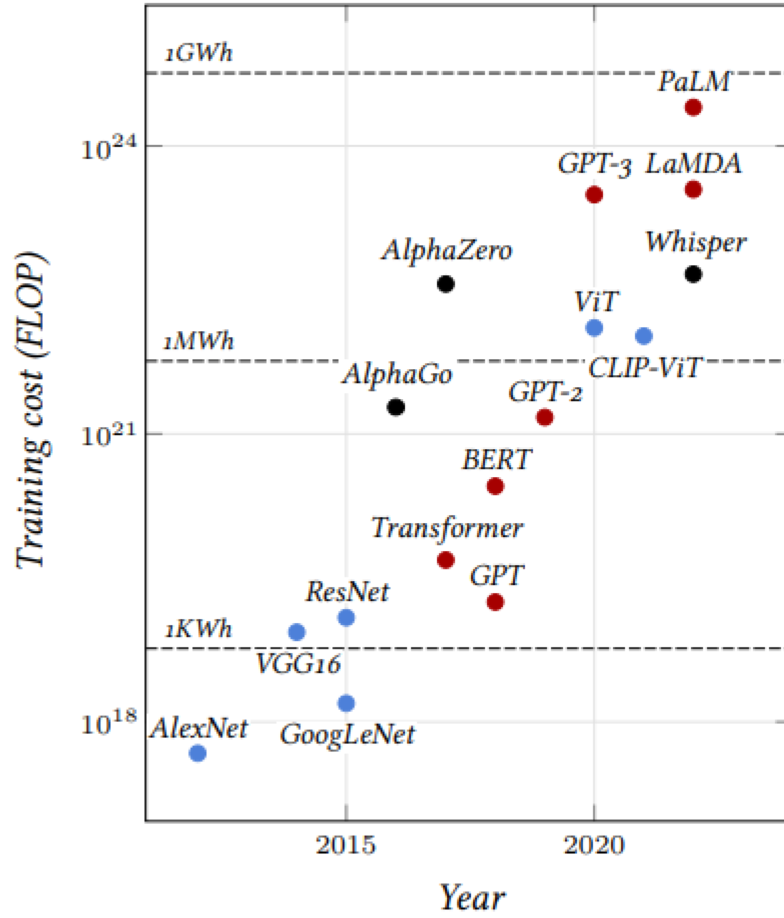






مخطأة الاختبار مع كمية الحوسبة في بيتا فلوب/باليوم (بيتا فلوب فلوب يعني حساب رقمي عائم وبيتا من المضاعفات العشرية 10^{15} يعني هلفًا). وحجم القاعدة بالكُليمات وحجم النموذج بالمتغيرات (المعاملات).

مستفيدا من هذه القوانين في النماذج الجلفية (البليونية) بفضل هيكلية الشعصبونات العميقة والتي تُحَجَّمُ بسهولة وكما سنرى فبزيادة عدد الطبقات أو أبعاد المُدخلات (المميِّزات) ولكن أيضا فضلُ طبيعتهم المُتوزعة (المنوزعة) حوسبيا وللمشتق المتدرج الدفعي (مُحَزَّم) والذي يتطلب فقد جُزءا من البيانات في وقت واحد تُحسب بدلا من البيانات التي تجاوز قُدرة الذاكرة المستخدمة. أدى هذا لنمو أسي



تكاليف التدريب بالفلبّات لبعض النماذج والألوان تحدد مجالاها فالأزرق صوري والأحمر لغوي وغيرهما أسود.
وأما الخطوط المتقطعة فإشارة للطاقة المُستهلكةباستعمال كروت أ100 SXM بدقة 16 بتًّا. وللمقارنة فاستهلاك الولايات المتحدة في 2021 يساوي 3920 تيرا واط TWh

عادة النماذج الصورية أن تكون بين عشرة إلى مئة جلف معامل (مليون) وتتطلب 10^{18}-10^{19} فلبَ تدريب بينما اللغويات فتبدأ من 100 مليون إلى مئات المليارات وتطلب 10^{20}-10^{23} فلبًا وتحتاج آلات (اجهزة) بكروت جوِّيدة.

تدريب هذه النماذج ممكن باستعمال مدونات (بيانات) بقاعدة صحيحة موسمة مُكلفة ولا يُنتج منها إلا حجم محدود. وعليه فيتم ذلك ببيانات مسحوبة من الشبكة تضمّ لها وهذا يؤدي إلى تكوين مُعطيات عديدة الصيغ (نصية ومرئي وصوتي) تُستعمل في تدريب أحجام مُشرفة كبيرة.

حتى 2024 أقوى النماذج المسماة نماذج لغوية ضخمة[22](نلضات) والتي ستأتي في الفصول القادمة، المدربة على مُعطيات (بيانات) جسيمة فطحلة.

[22] لا بُدَّ من اصطلاح على الصفة فالضخم ثم العملاق ثم المهول ثم وقد يضاف لها العارم

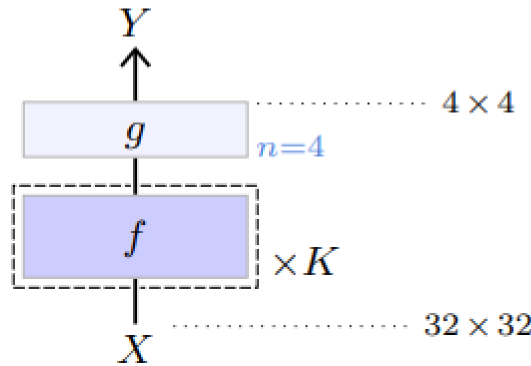
القسم الثاني النماذج العميقة

الفصل الرابع: مكونات النماذج

النموذج العميق مجرد تركيب رصّي (مرصوصي) حوسبي يحلل (يفكك) لعمليات رياضية أساسية من الجبر الخطي والتحليل. فعلى مد السنين طُورت أساليب عديدة (مجموعة) بمعانٍ وتعقيدات نمذجية مفيدة لمجالات عديدة. فالتجربة الرقمية ودليلها والنظرية ودليلها تظهر أداء أكبر يُبلغ بنماذج (بهياكل) أعمق وعليه فتركيب طويل من الإسقاطات كما مر معنا (3.4) وتدريب تلكم تحد بالنسبة لاضمحلال المتدرج وكثير من التحديثات التقنية انبثقت (ظهرت) من هذه المشكلة.

ترميز الطبقة (صوغ الطبقة)

نسمي الطبقات تركيبا أساسيا من العمليات المرصوصية (معقدة) صممت وجربت رقميا لتكون عامة (generic) وفعالة. أولئك عادة تحتوي متغيرات (وسيطات أو أوزان) مندربة وتشير إلى لمستوى تفاصيل لتصميم ووصف النماذج العميقة الضخمة. المصطلح موروث (مقتبس) من شعصبونة عديدة الطبقات، ومع أن النماذج الحديثة قد تتهيأ معقدة مكونة من عدة طرق (مسارات) متآزية.



في الصفحات التالية سأحرص على الالتزام بتركيب (بتصميم) النموذج أعلاه [للتسهيل]

- العمليات/الطبقات ممثلة كصناديق (مستطيلات)
- الغامقة تمثل (تتضمن) أوزان تدريبية
- القيم المنحكمة غير التلقائية مضافة على يمين مواضعها
- الغلاف (المحيط) المنقط (المخطط) يشير إلى تكرر هذه الطبقة عددًا من المرات كل واحدة بأزوانها.

- بعض الأحيان أبعاد المخرجات تُبين على اليمين عند عدم مساواتها للمدخلات

وإضافة الطبقات ذات البُنية الداخلية المعقدة تُصور بطول أكبر.

الطبقات الخطية [23]

أهم الوحدات في سياق الحوسبة وعدد المتغيرات (الأوزان) هي الطبقات الخطية فقد استفادت (واستفيدت) من عقود بحثية وهندسية في خرزمة الشرائح وتصيميها للعمليات المصفوفية.

ولاحظ أن المصطلح "خطي" في التعلم العميق يشير لعملية تآلفية مكونة من مجموعة تعبير خطي وانحياز ثابت.

طبقات كاملة الاتصال (متصلة بالكامل)

أبسط الطبقات الخطية هي مكتملة الاتصال موزونة بمندربات أوزان مصفوفية (مصفوفة أوزان) ﻭ لها حجم د×دَ ومتجه انزياح ب بعده دَ. هذه الطبقة تشكل تحويلا تآلفيا يُعمم (يوسع) لأشكال مرصوصية اعتباطية؟ حيث الأبعاد الزائدة تكون متجها دليليا (دلاليا). وصوغ (صياغة) ذلك بمعطًى (مدخل) س أبعاده $د_1× ... ×د_ك × د$ يحسب مخرجا ص أبعاده $د_1× ... ×د_ك × دَ$ مع

$\forall ل_1, ... ل_ك$

$$ص[ل_1, ... ل_ك] = ﻭس[ل_1, ... ل_ك] + ب$$

وفي الوهلة الأولى يُظن أن العملية التآلفية هذه تبدو محدودة بالتحويلات الهندسية مثل الدوران والتناظر والانزياح ولكنها قادرة على فوق (أكثر من) ذلك. تحديدا الاسقاطات لبعد مختزل أو إشارة مصفاة (تحديدا اسقاطات أبعاد مختزلة أو إشاراة مصفاة) أيضًا، فمن جهة كون الضرب الداخلي مقياسا للتشابه فضرب المصفوفة بالمتجه يُفهم كعملية تشابه رقمي بين المستعلمات المشفرة في متجه المدخل والمفاتيح المشفرة في صفوف المصفوفة.

كما مر معنا 3.3 فالمتدرج يبدأ بأوزان متكونة عشوائيا فإن فُعل هذا سذجيا كما في 3.4 فالشبكة قد تعاني من انفجار أو اضمحلال المفعلات والمتجهات []. الشعصبونات العميقة توفر (تنفذ) طرق ابتدائية مخصصة عند

---

[23]

شومءاب ج د و ح ط ي ك ل م ن س ع ف ص ق ش ت ث خ ذ ض غ ٮ ڡ ٯ ٮ ج ه ح ك ل ن س ع ص ق ت ث خ غ ج ح
ي ل ن س ص ق ش خ ض غ ں ڡ باجاحاطايا كامانا عافاصاقاشاتاخاضاظاغاڡاٮ ٮ ه هو ن ط ي ل م ن ع طايا كاناسا
عاصاقاشاتاثاخاظاغاڡاٮ ج ه هو ح ط ي ل م ن س ع ڡ ق ں ش ت خ ذ ض ظ غ ب ج د و ح ط ي ل م ن س ف ص ق
ش ت ث خ ذ ض ظ غ حـ د حـ د

التحجيم أن تجعل القيم (الأوزان) العشوائية طبقا لبعد المدخل ذات تباين كالمفعّلات ثابت وتجاوز (تمنع) الأداء الهش.

الشبكات التلافيفية

الطبقة الخطية تستقبل مدخلا بحجم اعتباطي (مرصوصًا) بتحويله لمتجه طالما أنه عنده العدد الصحيح من المعاملات. ولكن هذه الحلاة فقيرة (ضعيفة) التعامل مع المرصوصات الضخمة وطالما عدد المتغيرات وعدد العمليات مطردة لمضروب المدخل والمخرج أبعادهما. مثالُ ذلك لمعالجة صورة بفضاء لوني تحزي RGB بعدها 256×256 كمدخل ومخرج بنفس الحجم، فسيتطلب 4×10^{10} متغيرًا ومضاربات (عمليات ضرب).

بغض النظر عن هذه المشاكل العملية فإن معظم الإشارات ذات الأبعاد الكثيرة مهيكلة بقوة. مثال ذلك الصور تحتوي علاقات قصيرة المسافة (....) وثباتًا احصائية بالنسبة للترجمة، والحجم وبضع تناظرات وهذا لا ينعكس في الانحياز المستقرئ لطبقة كاملة الاتصال والذي يتجاهل بنية (هيكلة) الإشارة.

ولحل هذه المشاكل فالمختار استعمال الطبقات التلافيفية وهي أيضًا تآلفية تعالج البيانات (ذات البعدين) محليا بنفس العملية كل مكان.

التلافف ذو البعد الواحد يعرف (يُشكل) بثلاث متغيرات منحكمة: حجم النواة ن، عدد القنوات المدخلة: ق وعدد مخرجاته قَ والمتغيرات المندربة و لتطبيق تآلفي ϕ(.؛و):ح $^{ق×ن}$ ⟵ح $^{قَ×1}$

يمكنه معالجة أي مرصوصة س بحجم ق×ت حيث ت >= ق وتطبق ϕ(.؛و) لكل مُرَيصِصَة (جزء مرصوصة) بحجم ق×ن لـ س حافظةً النتائج في مصفوفة ص حجمها قَ×(ت-ق+1) كما في الصورة التالية (يسار). [اعتبرها تصغيرًا يضغط المصفوفة في متجه].

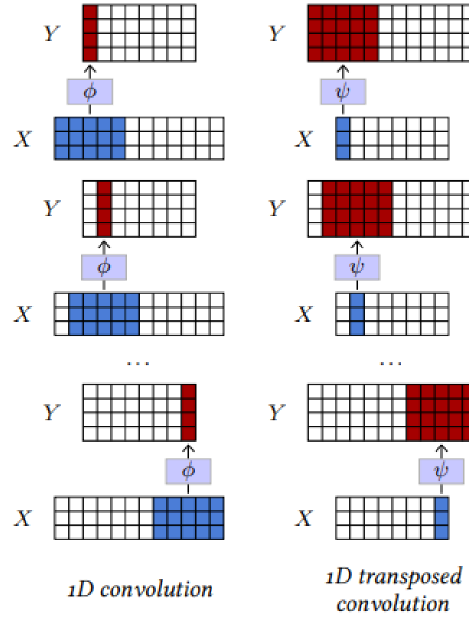


Figure 4.1: A 1D convolution (left) takes as input a $D \times T$ tensor X, applies the same affine mapping $\phi(\cdot; w)$ to every sub-tensor of shape $D \times K$, and stores the resulting $D' \times 1$ tensors into Y. A 1D transposed convolution (right) takes as input a $D \times T$ tensor, applies the same affine mapping $\psi(\cdot; w)$ to every sub-tensor of shape $D \times 1$, and sums the shifted resulting $D' \times K$ tensors. Both can process inputs of different sizes.

وأما التلافيفية ذات البعدين مشابهة لها نواة ن×ل (بعدا النواة) ومدخلها ن×ط×ع (المرصوصة) انظر الصورة التالية يسار.

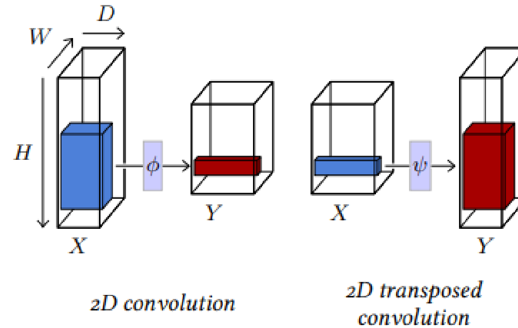


Figure 4.2: A 2D convolution (left) takes as input a $D \times H \times W$ tensor X, applies the same affine mapping $\phi(\cdot; w)$ to every sub-tensor of shape $D \times K \times L$, and stores the resulting $D' \times 1 \times 1$ tensors into Y. A 2D transposed convolution (right) takes as input a $D \times H \times W$ tensor, applies the same affine mapping $\psi(\cdot; w)$ to every $D \times 1 \times 1$ sub-tensor, and sums the shifted resulting $D' \times K \times L$ tensors into Y.

كلا العمليتين للمتغيرات المدربة لـϕ يمكن تصورها كمصفيات قَ بحجم ق×ن أو ق×ن×ل على التتابع ومتجه انحيازي بعده قَ.

طبقة كهذه مكافئة للترجمة (كأنها مترجمة) ويعني أن إشارة المدخل تُترجم فالمخرج يحول مشابهًا. وهذه الميزة تنتج انحيازًا استقرائيا مرادًا عند التعامل مع إشارة توزيعها ثابت مع الترجمة.

وأيضا -في الصورة- التالية هناك ثلاث متغيرات منحكمة إضافية:

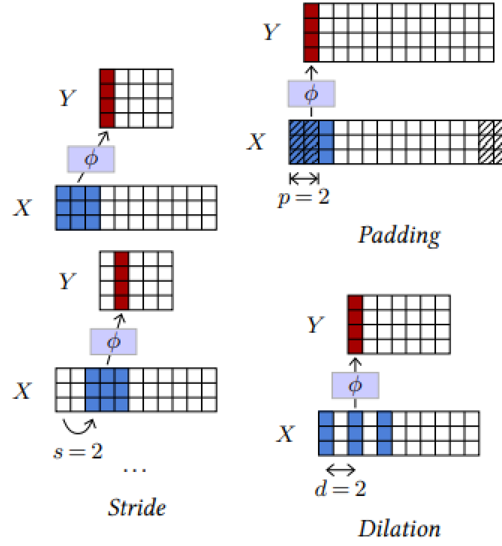


Figure 4.3: *Beside its kernel size and number of input / output channels, a convolution admits three hyper-parameters: the stride s (left) modulates the step size when going through the input tensor, the padding p (top right) specifies how many zero entries are added around the input tensor before processing it, and the dilation d (bottom right) parameterizes the index count between coefficients of the filter.*

(1) Padding الحشوة (البطانة، التوسيد): يحدد عدد الأصفار المضافة حول المرصوصة قبل معالجتها، لإبقاء حجم المرصوصة عندما تكون النواة أكبر من واحد فالقيمة تصير صفراً.

(2) Stride الفسح (الخطوة): تحدد قيمة الخطوة المستعملة عند المرور على المدخل تسمح لتقليل حجم المدخل هندسيا من خلال خطوات كبيرة والتلقائي واحد.

(3) Dilation التمديد تحدد عدد الدلائل بين معاملات المصفي للعملية التآلفية المحلية والتلقائي 1. والمكبرة منها تشير لإضافة أصفار بين المعاملات والذي يزيد حجم النواة/المصفي مع المحافظة عدد المتغيرات (الأوزان)

بغض النظر عن عدد القنوات فإن مخرج التآلفية عادة أصغر من مدخلها. فمثلا في حالة ذات البعد الواحد بدون الحشوة والتمديد فمدخل حجم ت ونواة حجمها ك والفسح س فحجم المخرج يصير:

تَ = (ت-ك)/س+1

مع تفعيل يحسب في طبقة التلفيف أو متجه بقيم لكل القنوات في موقع محدد فنسبة إشارة المدخل المعتمد عليها حقل استقبالي انظر الصورة التالية. على الطول والعرض ط×ع للمرصوص الجزئي المشير لإشارة واحدة د×ط×ع مرصوصة تفعيل تسمى: تطبيقا تفعيليا.

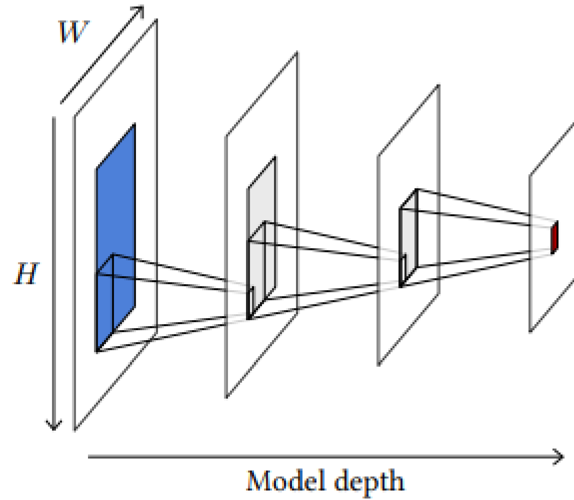


Figure 4.4: *Given an activation in a series of convolution layers, here in red, its receptive field is the area in the input signal, in blue, that modulates its value. Each intermediate convolutional layer increases the width and height of that area by roughly those of the kernel.*

التلافقيات يستعملون أيضًا لإعادة تركيب المعلومات عادة لتقليل الحجم الضخم للتمثيل في المقابل لعدد كبير من القنوات والتي تترجم إلى تمثيل محلي كثير (غني) يمكنها تطبيق عمليات اشتقاقية مثل محددات الأطراف أو تمثيل نموذج مطابقة. ونجاح مثل هذه الطبقة يمكن أن تصور كتمثيل تركيبي أو هرمي أو كعملية انتشار والذي المعلومات يمكن أن تُنقل بنصف حجم النواة المرور فيها.

هذه العملية الأخرى هي التلافف المنتقل أيضا تحتوي على عملية تآلفية محلية تعرف أيضا بمتغيرات منحكة ومندربة كالتلفف والتي مثالًا في حالة البعد الواحد يطبق Ψ(٠؛و):ح $^{قَ\times 1}$ ← ح $^{ق\times ن}$.

وعملية كهذه تزيد حجم الغشارة ويمكن فهمها كعملية تصنيع (وانظر الصور السابق الجانب الأيمن). وسلسلة الطبقات التلافقية هي نفس الهياكل التطبيقية لإشارة كبيرة البعد فمثل الصورة والصوت لمرصوصة صغيرة البعد. يمكن استعمالها للحصول على قيم الفئة عند التصنيف أو التمثيل المضغوط. طبقات التلافف المنتقلة تستعمل في

عملية عكسية لبناء أبعاد كبيرة من التمثيلات المضغوطة للوصول إلى التمثيلات المضغوطة الحاوية بيانات كافية لإعادة بناء الإشارة، وهو أسهل تعلم نموذج مكدوس (كثيف) على بعد صغير، وسيأتي أخرى.

الطبقات الخطية وحدها لا تساعد في تعلم العلاقات غير الخطية وذلك يضطرنا لعمليات غير خطية فهي تبرَمج (تبنى) مع دوال التفعيل والتي تحول (تطبق) كل عنصر من المرصوصة إلى رقم محدد منتجةً مرصوصة بنفس الحجم.

تتعدد دوال التفعيل وأشهرها ReLU (وحدة خطية تصحيحية) رَلّية (أو ريلو بالإمالة) والتي تُصَفِّرُ المُدخل السالب وتبقي الموجب كما هو:

$$\text{ريلو}(\text{س}) = \begin{cases} \text{٠,} & \text{س} > \text{٠} \\ \text{س} & \text{في الباقي} \end{cases}$$

وحيثُ أن تدريب النماذج العميقة معتمد على التدرج فُمشكلٌ أن يُختار تطبيق غير منشق عند الصفر وثابت شطر الخط. ومع ذلك فإن المتدرج المنحدر مراده الأهم أن يكون تدرُّجه ناجعًا (مفيدًا) في المتوسط. فتهيئة المتغير وتسوية البيانات تجعل نصف المُنَشِّطات موجبة عند بدء التدريب كما يُراد.

قبل تعميم الرلية كان المشهور هو الظلّ الزائد (ظاز/ظلز) (انظر الصورة التالية) والذي يتشبع أسيّا بسرعة على الجانبين الموجب والسالب مفاقمًا اضمحلال التدرج.

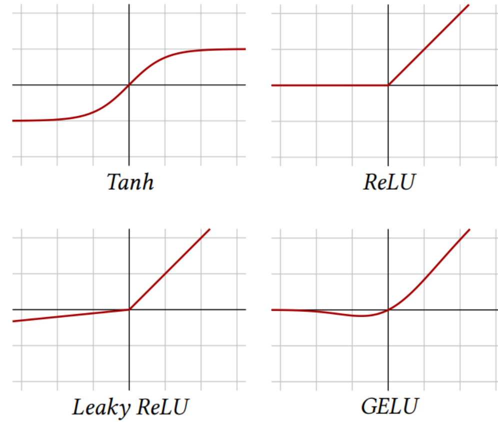


Figure 4.5: *Activation functions.*

دوال تنشيط أخرى مشهورة تتبع نفس الفكرة المحافظة على الموجبات غير مُغَيّرة (ملموسة) بينما تهرس/تسحق السوالب. الرلية السربة تجدو السوالب بقيمة موجبة صغيرة (انظر الصورة التالية)

$$\text{leaky relu}(x) = \begin{cases} ax \text{ if } x < 0, \\ x \text{ otherwise.} \end{cases}$$

وأما جيلو (وطمية) فمُعَرَّفةٌ باستعمال دالة التوزيع التراكمي للتوزيع الغوسيِّ كالتالي:

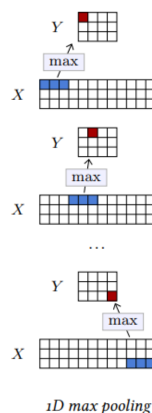
$$\text{gelu}(x) = xP(Z \leq x),$$

حيث z من التوزيع الطبيعية تط(0،1) وهي تسلك نحوَ الرلية السلسة (انظر الصورة الرباعية) اختيار دالة التفعيل/التنشيط عبر تشكيلات الرلية أمر عادته التجربة التطبيقية.

التجميع Pooling

الخطة التقليدية لتقليل حجم الإشارة هي استعمال عملية التجميع الدامجة لعدة منشطات في واحدة ملخصة المعلومة. أشهرها المعيارية هي عملية التجميع المُتستعظم max وهي شبيهة بالترشيح تعمل على بعد أو اثنين معرفة بحجم نواة. في هيئتها المعيارية هذه الطبقة تحسب أعظمية المنشطة لكل قناة فوق رُصيصات (sub-tensors) متداخلة الحجم المكاني مساوٍ لحجم النواة. هذه القيم تُحفظ في مرصوص ناتج فيه نفس عدد القنوات كالمدخل واللاتي حجمها المكاني مقسم بحجم النواة. وكما الترشيح هذه العملية لها ثلاث متغيرات مُنحكمة: التلبيدة (الغلقة)، الخطوة، التمدد padding, stride, and dilation، حيث الخطوة تساوي حجم النواة عادةً (تلقائيا) بينما الخطوة الصغيرة تنتج مرصوصة أكبر كما في معادلة الترشيح السابقة (رسمة المصفوفات الملونة مع الأنوية).

عملية الاستعظام يُمكن أن تفهم وتترجم لإقصاء منطقي حيثُ أنها تتبع سلسلةً من الطبقات المرشحة والتي تحسب قيمًا محلية للقطع الحاضرة لتشفير السابقة كقطعة من اللاحقة واحدةٍ على الأقل. فهي تخسر دقة المكان جاعلةً بديلها المعياري هو التجميع الموَسّط إذ طبقته تحسب المتوسط بدلا من استعظام الرُّصيصات وهذه عملية خطية عكس التجميع المستعظم.



1D max pooling

Figure 4.6: *A 1D max pooling takes as input a $D \times T$ tensor X, computes the max over non-overlapping $1 \times L$ sub-tensors (in blue) and stores the resulting values (in red) in a $D \times (T/L)$ tensor Y.*

الفجونة (الإسقاط)

بعض الطبقات تصمم لهدف محدد يسهل التدريب أو تحسين التمثيلات (المتجهات التضمينية المختزلة).

أحد هذه الاسهامات الأساسية من هذا هو <u>الفجونة</u> فمثل هذه لا متغيرات مندربة فيها بل متغير منحكم واحد p تستقبل مرصوصة بأي حجم.

هي عادة تُطفأ أو لا تشغل أثناء الاختبار وهي التي مخرجها كمدخلها أما عند تفعيلها فعندها احتمال p يُصَفِّرُ بعض مفعلات المدخل مُستقلة، ثم تُحَجِّمُ كُلَّ المفعلات بعامل p-1/1 لإبقاء القيمة المتوقعة غير متغيرة (انظر الرسمة التالية):

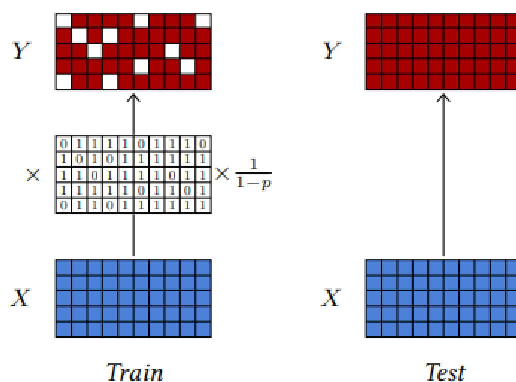


Figure 4.7: *Dropout can process a tensor of arbitrary shape. During training (left), it sets activations at random to zero with probability p and applies a multiplying factor to keep the expected values unchanged. During test (right), it keeps all the activations unchanged.*

الدافع وراء الفجونة/الفرغنة تفضيلُ التنشيطات الفردية بدلا (بكف أو تثبيط) التمثيل المجموعي. فبما أن الاحتمال لمجموعة تنشيطات k الباقية سليمة عبر طبقة الاسقاط هو $(1-p)^k$ التمثيلات المتداخلة تُصبح غير معتمدٍ عليها تجعل المنهجية التدريبية تتفاداهم (تتجاهلهم) وأيضًا قد ترى حقنًا ضوضائيا تُثَبِّتُ التدريب أكثرَ.

عند التعامل مع الصورة والمرصوصات ثنائية الأبعاد فالترابط قصير المدى للإشارات والتكرار الناتج يمنع تأثير الفجونة لأن التنشيطات المُصَفَّرة قد تَصفُر بسبب جيرانها وعليه ففرغنة مرصوصات ثنائية الأبعاد تصفر قنوات كاملة بدلا من منشطات فردية (انظر الصورة التالية)

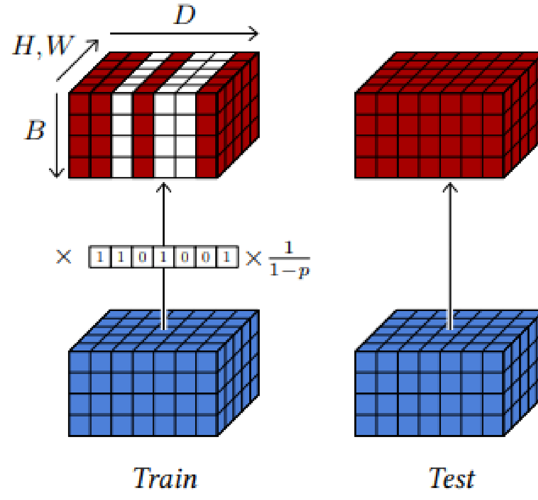


Figure 4.8: *2D signals such as images generally exhibit strong short-term correlation and individual activations can be inferred from their neighbors. This redundancy nullifies the effect of the standard unstructured dropout, so the usual dropout layer for 2D tensors drops entire channels instead of individual values.*

ومع أن الفجونة مستعملة لتحسين التدريب غير عاملة (مفعلة) في التشغيل ويُمكن استعمالها في اعدادات محددة كخطة استعشائية مثل أن تقيّم درجات الثقة التجريبية. empirically confidence scores.

طبقات التسوية (النظمنة)

قسم مهم من العمليات المساعدة (المساعدة) في الهيكليات العميقة هي <u>طبقات التسوية</u> والتي تغصبُ (ترغم) الوسيط التجريبي وتباعد مجموعات المنشطات.
الطبقة الأساسية في هذه العائلة هي <u>مُسَوِّيَةُ الحزمة</u> وهي الطبقة المعيارية لمعالجة الحُزَم بدلا من العينات الفردية.
بمتغير منحكم D وسلسلتين من السلميات المندربة بيتاوي1، ... بيتاوي$_D$ وكذا الغاميَّ مثلها.
فمثلًا حُزمة معطاة فيها B عينات س1، ... س$_B$ ببعد D فهي تحسب لكل بُعدٍ منها متوسطا وتباعدها عبرَ الحُزمة:

$$\hat{m}_d = \frac{1}{B}\sum_{b=1}^{B} x_{b,d}$$
$$\hat{v}_d = \frac{1}{B}\sum_{b=1}^{B} (x_{b,d} - \hat{m}_d)^2,$$

ولكل منها تُحسب لكل قطعة $x_{b,d}$ تُحسب قيمة مسَوَّاة z بمتوسّط صفر وتباعد 1، ومنها قيمة نهائية $y_{b,d}$ مع متوسّط بيتاوي$_d$ وتباعد غامّوي$_d$

$$\forall b, \quad z_{b,d} = \frac{x_{b,d} - \hat{m}_d}{\sqrt{\hat{v}_d + \epsilon}}$$
$$y_{b,d} = \gamma_d z_{b,d} + \beta_d.$$

ولأن هذه التسوية معرفة عبرَ الحزمة فهي مُستعملة (أو مقامٌ بها) في التدريب فقط أما الاختبار في الاختبار فالطبقة تحول كُل عينة منفردةً بناءً على المتوسط والتباعد المحسوبين على كل مجموعة التدريب، والتي تنزل تحويلًا تآلفيًّا لكل قطعة.

الدافع وراء تسوية الحُزمة كان لتجنب التغير في الحوجمة عبر طبقة الشبكة عند التدريب فتؤثر على كل الطبقات التالية (اللاحقة) والتي سوف تحصِّلُ المتغيرات المندربة تتابعيًّا. ومع أن فنط العمل قد يتعقد فوق هذا الدافع، فهذه الطبقة تحسّن تدريب النماذج العميقة.

أما في حالة المرصوصات ثنائية الأبعاد تتبع مبدأ طبقات الترشيح لتعالج كل الأماكن متشابهةً فالتسوية تكون عبر كل قناة عبر كل الأماكن ثنائية الأبعاد والبيتاويّ والغاميّة تبقى متجهات D فالتحجيم والنقل (المناوبة) لن تعتمد على المكان البُعداني وعليه إن كان المرصوص المعالج له أربع أبعاد: B ×D ×H ×W فالطبقة تحسب (md,vd) لكل D ... ,d=1 من القطعة: B ×H ×W فتُسَوَّى حسبهم ثم تحجم وتزاح (تُنقل) قيمها (مكوناتها) بقيمة المتغيرين المندربين البيتاوي والغامّوي βd and γd.

وعليه فمرصوصة مُعطاة B ×D تسويةُ الحزمة تُسَوّيها عبرَ b وتُقيّسُها/وتزيحها موضعها d الذي يُنَجّز كقطعة تُجدى قطعيًّا بـ γ تُجمع مع β. فعندما تكون أبعاده: B ×D ×H ×W فهي تُسَوّي عبر b,h,w وتقيّس وتزيح طبقًا لـ d انظر الصورة القسم اليسار

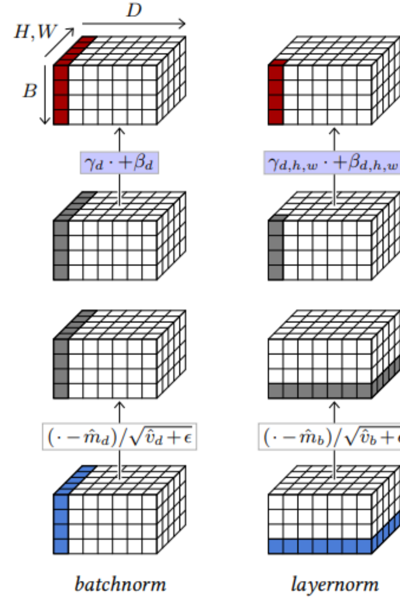


Figure 4.9: *Batch normalization (left) normalizes in mean and variance each group of activations for a given d, and scales/shifts that same group of activation with learned parameters for each d. Layer normalization (right) normalizes each group of activations for a certain b, and scales/shifts each group of activations for a given d, h, w with learned parameters indexed by the same.*

وهذا يُمكن تعميمه بناءً على هذه الأبعاد، مثلًا طبقة التسوية تحسب أزخام وتسويات كل المركبات لكل عينة مفردةٍ منها، وتقيّس وتزيح المركبات أفرادًا (انظر الصورة السابقة اليمين) طبقًا له أيضًا. فالمرصوص B ×D ×H ×W يُسوَّى عبر d,h,w ويقيس ويزاح طبقًا لها أيضًا.

عكسًا لتسوية الحُزمة والتي تعالج العيّنات مُفردَةً فطبقة التسوية تتصرف بشكل متشابه خلالَ التدريب والاختبار.

الوصلات المتجاوزة (المتخطِّية)

أسلوبٌ آخرُ يُخفِّفُ اضمحلال التدرج ويسمح (يدعم) تدريب هياكل عميقة يُسمى وصلات التجاوز (التخطّي) وهؤلاء ليسوا طبقات بل تصميم هيكلي مُخرج بعض الطبقات يُنقل كما هو لأخرى عبرَ النموذج، متجاوزًا المعالجة (الجداءات) بينها. هذه الإشارة غير المغيّرة (المعالجة) قد تضافا أو تحاذيا (تُسَلسلا) concatonate مع مدخل الطبقة الفروع منها.

فائدتها المرادة والمقصود من هذا التصميم هو التأكد حتى وإن قُتل التدرج في حالةٍ (وضعية) ما فالتدرج سينتشر عبر الوصلات المجاوزة.

الوصلات الفضالية Residual (المتبقيّة) تحديدًا تعين وتسمح ببناء شبكات بمئات الطبقات وبعض النماذج مثل نماذج الفضالية في حوسبة الصور (الحسوبة الصورية) والمحولات في معالجة اللغات الطبيعية مركبة كليا من كّل من طبقات الوصلات الفضالية.

ودورها أيضا أن تساعد التفسير عديد الأدوار في النماذج التي تقلل حجم الإشارة قبل إعادة تكبيرها عبرَ توصيل طبقات بالأحجام المتناسبة (انظر الصورة):

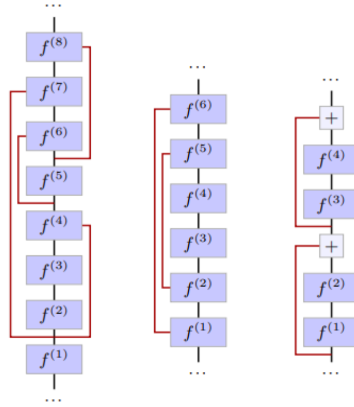


Figure 4.10: *Skip connections, highlighted in red on this figure, transport the signal unchanged across multiple layers. Some architectures (center) that downscale and re-upscale the representation size to operate at multiple scales, have skip connections to feed outputs from the early parts of the network to later layers operating at the same scales [Long et al., 2014; Ronneberger et al., 2015]. The residual connections (right) are a special type of skip connections that sum the original signal to the transformed one, and usually bypass at most a handful of layers [He et al., 2015].*

مثاله التجزئة المعنوية (التقطيع المعنوي) في حالة الوصلات الفضالية فقد تساعد التعلم بتبسيط (تسهيل أو تيسير) مهمة إيجاد تحسينات جُزئية بدلًا من تحديثٍ كامل.

طبقات الإنتباه

في عدة تطبيقات يُحتاج إلى عملية تدمج المعلومات المحلية لأماكن بعيدة في المرصوصة. مثالُ ذلك التفاصيل الدقيقة لصورة واقعية دقيقة مولدة أو كلمات في مواضع (أماكن) مختلفة في الفقرة طلبًا لقرار نحوي أو معنوي في معالجة اللغات الطبيعية.

فالطبقات كاملة الاتصالات لا تستطيع معالجة الإشارات كبيرة الأبعاد، ولا الإشارات متغيرة الحجم ولا طبقات الترشيح تستطيع نشر المعلومات بسرعة. الطرق التي تدمج نتائج المرشحات عبرَ توسيطها في مساحات مكانية (مسحية) كبيرة تعاني من خلطِ عدة إشارات لعدة أبعاد محدودة.

طبقات الانتباه تختص بحل هذه المشكلة عبر حساب قيمة انتباه لكل وحدة في مرصوصة النتيجة لكلٍّ في المرصوصة المُدخلة بدون قيود محليّة ومع توسيط السمات عبر المرصوص كُلُّه تتابعًا.

ومع أنها أعقد من الطبقات الأخرى، فقد أصبحت عُنصُرًا معياريًّا في عدد من النماذج الحديثة. فمثلًا هي المُرَكَّب الأساسي في المحوِّلات الهيكليّة المُسيطرة (المهيمنة) للنماذج اللغوية الضخمة (وسيأتي).

عامل الانتباه

مُعطًى

- مرصوص Q من المُستعلِمات حجمُها $N^Q \times D^{QK}$
- مرصوص K من المفاتيح حجمُها $N^{KV} \times D^{QK}$
- مرصوص V من القِيَم حجمُها $N^{KV} \times D^V$

وعملية الانتباه تَحسِبُ مرصوصًا

$$Y = \text{att}(Q, K, V)$$

بُعدها $N^Q \times D^V$ ولِنَصِل لذلك نحسبُ أوَّلًا لكل مُستعلمة الخانة (المؤشر/دليل) q ومِفتاح دليله k قيمة انتباه $A_{q,k}$ كمُستعظم مكاني ناعم (محصي مُستعظم) للجداءات الداخلية بين المُستعلِم Q_q والمفاتيح:

$$A_{q,k} = \frac{\exp\left(\frac{1}{\sqrt{D^{QK}}} Q_q \cdot K_k\right)}{\sum_l \exp\left(\frac{1}{\sqrt{D^{QK}}} Q_q \cdot K_l\right)}$$

حيثُ معامل التحجيم المحافظ على $\frac{1}{\sqrt{D^{QK}}}$ مدى القِيَم تقريبًا غير متغيّرة ولو لكبار D^{QK}.

ثم القيمة المُسترجعة تُحسب بتوسيط القيم بناءً على قيم الانتباه:

$$Y_q = \sum_k A_{q,k} V_k.$$

ثم إن كانت القيمة Q_n توافق (تماثل) مفتاحًا K_m أكثر من غيرها فقيمة الانتباه المقابل $A_{n,m}$ ستقارب الواحد، والقيمة المُسترجعة Y_n ستكون القيمة V_m المرتبطة بذلك المفتاح.

وإن كان المستعلم Q_n يوافقُ مفتاحًا K_m أكثر من غيره، فقيمة الانتباه المشير لهما $A_{n,m}$ سيكون قريبا من الواحد والقيمة المسترجعة Y_n ستكون القيمة V_m المرتبطة بالمفتاح، ولكن إن شابهت عدة مفاتيح بتساوٍ فإن Y_n ستكون متوسط القيم المتعلقة وهذا يُمكن صوغُه كالتالي:

$$\operatorname{att}(Q,K,V) = \underbrace{\operatorname{softargmax}\left(\frac{QK^{\top}}{\sqrt{D^{QK}}}\right)}_{A} V.$$

هذه العملية عادة تُوسَّع بطريقتين كما في الشكل التالي،

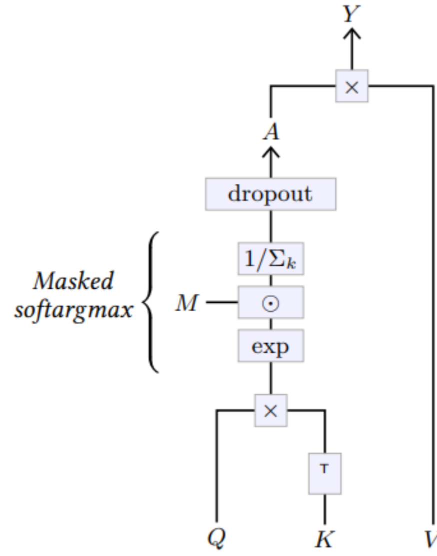


Figure 4.12: *The attention operator $Y = \operatorname{att}(Q,K,V)$ computes first an attention matrix A as the per-query softargmax of $QK^{\top}$, which may be masked by a constant matrix M before the normalization. This attention matrix goes through a dropout layer before being multiplied by V to get the resulting Y. This operator can be made causal by taking M full of 1s below the diagonal and zeros above.*

أولا مصفوفة الانتباه تقنّع بضربها قبل تنظيم استعظام المحصي بمصفوفة بُلْيانيّة (إثنانية) M، والتي تسمح بجعل العملية سببية بجعل M مليئة بـ1 واحدات تحت قطرها وأصفارًا فوقه مُجنّبةً Y_q من رؤية المفاتيح والقيم لمؤشرات (دلائل) k أكبر من q والثاني مصفوفة الانتباه تعالج بطبقة مسقطة (مثغرة، مفجوة، ثغرّة ...) (انظر الأقسام السابقة) قبل ضربها بـV مُفِيدَتَنْهَا بفوائدها.

وحيثُ يحسب الجداء الداخلي لكل زوج مفتاح ومستعلم فالكلفة الحسابية لعملية الانتباه مكعبة مع طول السياق، وهذا بان اشكال إذ بعد تطبيقات هذه الأساليب تتطلب معالجة عشرات آلاف السياقات (السلاسل) أو أكثر. هناك عدة محاولات لتقليل هذه الكلفة مثل دمج انتباهًا كثيفًا لنافذة محلية مع انتباه متباعد طويل المدى[24] أو باستخطاط (تخطيط أو جعلِ شيء خطيًّا) العملية للساتفادة من تجميعية (تبادلية، اقترانية) مضروب المصفوفة وحساب المفتاح والقيمة قبل ضربها بالمُستعلمات[25]

ملاحظة: قم بوضع المصادر في الحاشية كما هنا لتستخدمها عند شرح الكتاب.

[24] Longformer: The Long-Document Transformer

[25] Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention

طبقة الأنيبهات (الانتباه عديد الرؤوس)

عملية الانتباه عديمة المتحكّمات (المتغيرات) هو المفتاح الرئيسي في <u>طبقة الانتباه عديد الرؤوس</u> ممثلة في الرسم

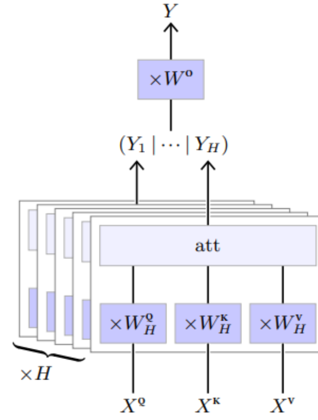


Figure 4.13: *The Multi-head Attention layer applies for each of its $h = 1, \ldots, H$ heads a parametrized linear transformation to individual elements of the input sequences X^Q, X^K, X^V to get sequences Q, K, V that are processed by the attention operator to compute Y_h. These H sequences are concatenated along features, and individual elements are passed through one last linear operator to get the final result sequence Y.*

هيكلية (بُنية) هذه الطبقة تعرف بعدة <u>متغيرات منحكمة</u>: عدد الرؤوس H وأحجام ثلاث مصفوفات مندربة متستسل في H:

- W^Q of size $H \times D \times D^{QK}$,
  W^K of size $H \times D \times D^{QK}$, and
  W^V of size $H \times D \times D^V$,

ولحساب المستعلمات والمفاتيح والقيم من المُدخل تتابعيا والقيمة النهائية للمصفوفة W^O ذات الحجم $HD^V \times D$ لتراكم (تكّل، تكدّس، تكلل ...) نتائج بالنسبة لكل رأس (من كلّ رأس).

هي تأكد مُدخلها بثلاث سلاسل (متتابعات):

- X^{Q} of size $N^{\text{Q}} \times D$,
- X^{K} of size $N^{\text{KV}} \times D$, and
- X^{V} of size $N^{\text{KV}} \times D$,

ومن كل واحدة تحسب لكل: $h = 1, \ldots, H$,

$$Y_h = \text{att}\left(X^{\text{Q}} W_h^{\text{Q}}, X^{\text{K}} W_h^{\text{K}}, X^{\text{V}} W_h^{\text{V}}\right).$$

هذه المتتابعات (السلاسل) $Y_1, \ldots, Y_H$ تلُمّ (تضم) عبرَ بُعد الميزة (السمة) وكل واحد من النواتج يُضرب بـ W^{O} للحصول على آخر النتائج:

$$Y = (Y_1 \mid \cdots \mid Y_H) W^{\text{O}}.$$

وكما سنرى في الفصول التالية ان شاء الله هذه الطبقة تستخدم لبناء بِنية جزئية لنموذجين:

قوالب الانتباه الذاتي (كل، تجميعة) والتي فيها ثلاث سلاسل (متتابعات، مترادفات، متعاقبات) X^{Q}, X^{K}, and X^{V} هم نفس بعضهم أو قوالب الانتباه المتقاطع حيثُ الثانية لا تساوي الثالثة

تضمين الكُليمة

في عدة حالات نحتاج تحويل الكُليمات (الأجزاء، الوحدات، مصوغ) إلى متجهات، وهذا يمكن من خلال طبقة التضمين والتي تحتوي على جدول مُنبحث (منرمق) يصلُ مباشرة الصحيحات (الأعداد) بالمتجهات. طبقة كهذه تعرف بمتغيّرين منحكمين: N وهو عدد قيم الكُليمات الممكن وبُعد المتجهات المُخرجة D فتصير مصفوفة M مندربة أبعادها N×D.

مُعطًى مدخل رقمي (رقم صحيح) X أبعاده $D_1 \times \cdots \times D_K$ والقيم في $\{0,\ldots,N-1\}$ فهذه الطبقة ترجع مرصوصة حقيقية القيم Y أبعادها $D_1 \times \cdots \times D_K \times D$ مع التالي:

$$\forall d_1,\ldots,d_K,$$
$$Y[d_1,\ldots,d_K] = M[X[d_1,\ldots,d_K]].$$

تشفير الموقع

بينما عولجت الطبقة كاملة الاتصال محددة بالنسبة لمكان لميزات المرصوصة المُدخلة ومواقع التنشيطات الناتجة في المرصوصة المُخرجة، والطبقات التلافقية وطبقات الانتباه عديد الرؤوس فهي مُغفلة للموقع المُطلق في المرصوص، وهذا مفتاح لقويَّتهم غير المتباينة والانحياز المستقضي وهو مفيدٌ للاشارة الثابتة.

ولكن هذا قد يكون مشكلا في حالات محددة حيثُ المعالغة ينبغي أن تصل للموقع المطلق ومثالُ ذلك التصنيع الصوري حيثُ احصاءات المنظر ليست ثابتة أو في معالجة اللغة الطبيعية حيث المواقع النسبية للكلمات قوية التصويغ لمعنى الجملة.

الطريقة المعيارية للنسخ لهذه المشكلة هي أن تضيف أو تضم تمثيل الميَز في كل موقع وهو التشفير الموقعي (الممكنة، الموقعة) وهو متجه ميزي يعتمد على موقعٍ في المرصوصة، وهذا التشفير الموقعي يمكن أن يتعلَّم مثل غيره من متغيرات الطبقة أو أن تعرف تحليليًّا (رقميًّا).

مثاله: في نموذج المحوّل الأصلي فلكل متجهات متتابعة بُعدها D يضافُ لها تشفير لدلائل السلسلة كمترادفة (متتابعة) من الجيبيَّات (sines) والجيبُتميَّات (cosines) على ترددات مختلفة:

$$\text{pos-enc}[t,d] = \begin{cases} \sin\left(\frac{t}{T^{d/D}}\right) & \text{if } d \in 2\mathbb{N} \\ \cos\left(\frac{t}{T^{(d-1)/D}}\right) & \text{otherwise,} \end{cases}$$

with $T = 10^4$.

الفصل الخامس: البُنى (الهيكليات)

مجال التعلم العميق (التعامق) طور خلال السنين لكل تطبيق مجالي عدة هيكليات عميقة وهي توفر توازناً جيداً بين العديد من المعايير المهمة، مثل سهولة التدريب ودقة التنبؤ وحجم الذاكرة والتكلفة الحسابية وقابلية التوسع.

الشُّعصوبات عديدة الطبقات[26]

أبسطُ بُنية (هيكلية) في التعلم العميق الشُّعصوبات عديد الطبقات (MLP) أو الشُّعاصيبات ولها شكلٌ يخلفُ الطبقات كاملة الاتصال مُفَرَّقة بدوال التنشيط (التفعيل) انظر مثالًا الرسم التالي:

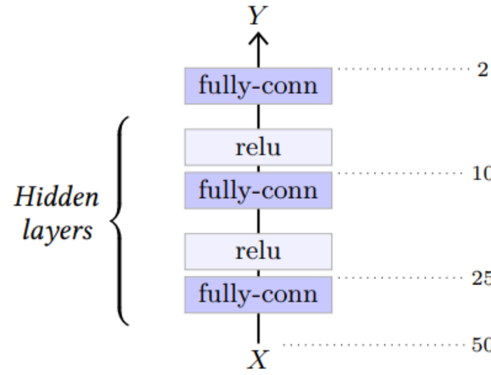


Figure 5.1: *This multi-layer perceptron takes as input a one-dimensional tensor of size 50, is composed of three fully connected layers with outputs of dimensions respectively 25, 10, and 2, the two first followed by ReLU layers.*

ولأسباب تاريخية في هكذا نموذج فعدد الطبقات المخفية يشير إلى عدد الطبقات الخطية بدون آخر واحدها.
نتيجة مفتاحيّة نظرية هي نظرية التقريب العامّة[27]. والتي تنصُّ على إن كانت دالة التنشيط σ متصلة (مستمر) وليست حدوديّة، فأي دالة مستمرَّة f يُمكن تقريبها تعسّفيًّا (اعتباطي) بشكل جيّد بانتظام على مجال متراصّ، وهو محدود ومتضمّن (حاوٍ) لحدوده، بنموذج بالصيغة (الهيئة أو تركيب) $l_1 \circ \sigma \circ l_2$ حيث الأول والثاني تآلُفيَّان ffine حيثُ النموذج شُعاصيبٍّ بطبقة مخفيّة واحدة، وهذه النتيجة تقتضي أنها تستطيع تقريب أن شيء ذي قيمة عملية، ولكن هذا التقريب اشتمل على بُعدٍ مخرج الطبقة الخطية الأولى يكون كبير اعتسافًا.
وببعدٍ عن يسورَتِه (بساطته) فالشُّعاصبيَّات تبقى أداة مهمة حيثُ بُعد الإشارة المعالجة ليست عملاقةً (ضخمة، عظيمة، هائل).

---

[26] كلمة Perceptron في معجم سدايا منقحرة وفي المعاني "عصبون مدرك" و "مستقبل متعدد الطبقات" وعندنا في الأعصاب عُصبُون وهو فُعلون والله أعلم من عصب وأنا وضعت شعصب نحتاً للشبكة العصبية فقد يصح منها شُعْصوب فعلول أو نجذّر لثلاثي فتصير شعص (وليس جذرًا مستعملاً) فتصير شُعْصون

[27] https://web.njit.edu/~usman/courses/cs677/10.1.1.441.7873.pdf

الشبكات التلاففية (تلافّية)

الهيكلية المعيارية لمعالجة الصور هي الشبكة التلاففية أو الشكفيَّات التي تدمج عدة طبقات تلاففية إما لتقليل حجم الإشارة قبل معالجها بالطبقات كاملة الاتصال لإخراج إشارة ثنائية الأبعاد أيضًا كبيرة الحجم.

شبيهة LeNet

النموذج LeNet الأصلية لتصنيف الصور يجمع طبقات تلاففية متتابعة ثنائية الأبعاد مع طبقات تجميع مُستعظم تلعبُ دورًا في استخراج الميَز مع سلسلة طبقات كاملة الاتصال شعاصبيَّةً وتعمل التصنيف لكل عيّنة.

هذه الهيكلية كانت حجر الأساس (مخطط) لعدة نماذج تتشارك هيكلها ولكنَّهم أكبر مثل: AlexNet وعائلة VGG

الشبكات المتجاوزة Residual

الشبكات التلاففية المعيارية التابع للعائلة السابقة ليس سهلًا توسيعها لهيكليات عميقة، فهي تعاني من مشكلة اضمحلال المشتق. الشبكات المتجاوزة (الشبجات) أو ما يسمى بـ ResNets تحدد تمامًا هذه المشكلة وتحلُّها بدقة مع الوصلات المتجاوزة والتي تسمح لمئات الطبقات. أصبحت هذه الهيلكية معيارية في هياكل الرؤية الحسوبية وتطبيقاتها وقد وسّعت في عدة نسخ بناءًا على عدد الطبقات، نحن سننظر مدققين في المصنف ResNet-50 وهيكليّته.

كما أخواته فهو مكوّنة من سلسلة قطع (كتلة، لبنة، قوالب) متجاوزة كل منها تحتوي عدة طبقات تلاففية وطبقات نظيم الدُّفعة وطبقات رليّة (ReLU) مغلفة في اتصال متجاوز وفي الصورة التالية تصوّر القطعة:

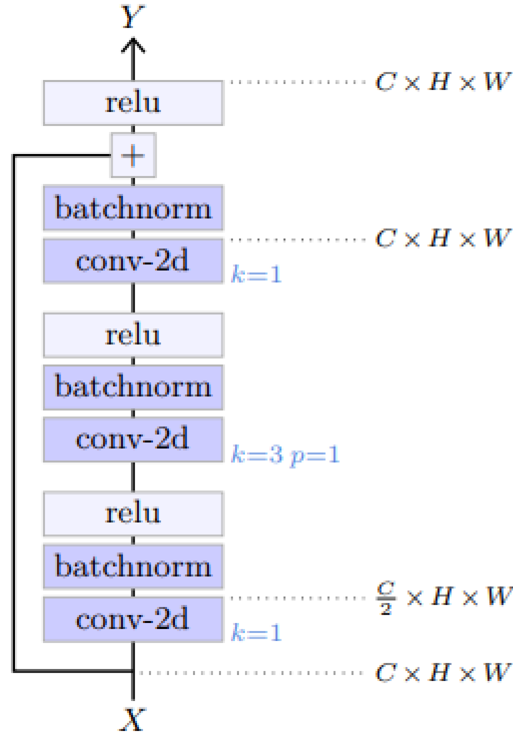


Figure 5.3: *A residual block.*

متطلب جوهري للأداء العالي مع الصور الحقيقية أن تُنشر الإشارة مع عدد كبير من القنوات للسماح لتمثيل غني، ولكن عدد متغيرات الطبقات التلافيفية مع تكلفتها الحوسبية مستكعبة مع عدد القنوات. هذه الطبقة المتجاوزة تحلحل هذه المشكلة أولا بتقديم عدة قنوات مع تلافف 1×1 ثم عمل تلافف 3×3 مكانيًّا على العدد مختزل القنوات ثم رفع عدد القنوات مرة أخرى مع تلافف 1×1.

الشبكة تقلل بُعد الاشارة أخيرا لحساب المُحُض للتصنيف. وهذا يتم شُكرًا للهيليكة المركبة من عدة أقسام كل منها تبدأ بـ قطعة متجاوزة مُصَغّرة downscaling والتي تنصّف طول وعرض الاشارة وتضاف عدد القنوات متبعة سلسلة قطع متجاوزة. وقطعة التصغير المتجاوزة لها هيكلية مشابهة للقطعة المتجاوزة المعيارية إلا أنها تتطلب اتصالا متجاوزًا يغير حجم المرصوص. وهذا يتم مع التلافف 1×1 مع مخطوة stride قيمتها اثنان (انظر الصورة التالية):

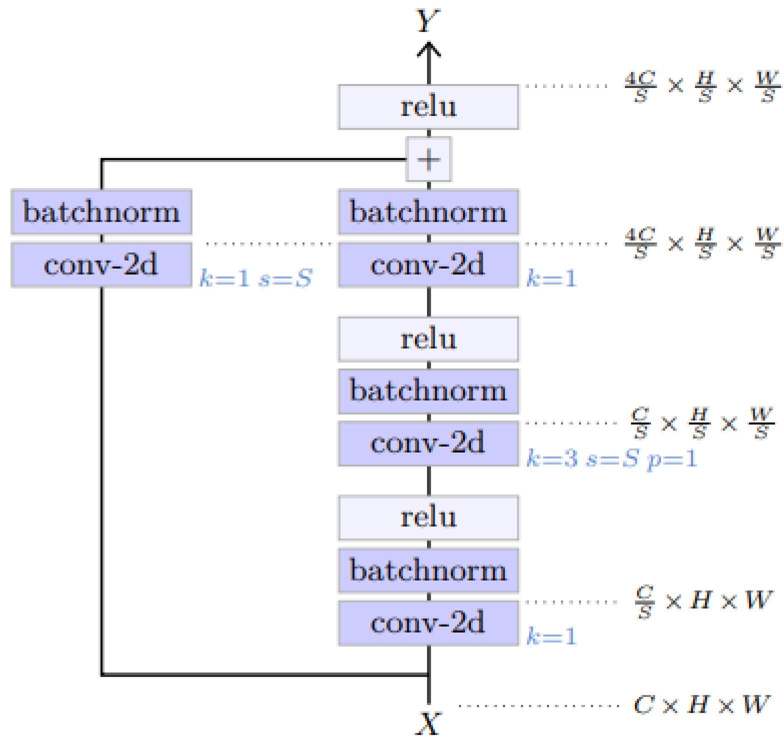


Figure 5.4: *A downscaling residual block. It admits a hyper-parameter S, the stride of the first convolution layer, which modulates the reduction of the tensor size.*

فالهيلكية العامة لـ ResNet-50 عُرضت في الشكل التالي[28]:

[28] المصدر: https://arxiv.org/pdf/1512.03385

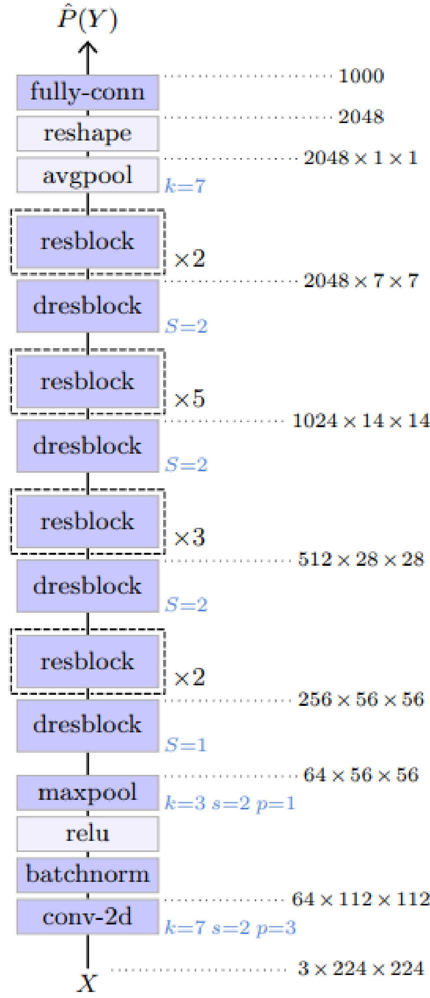


Figure 5.5: *Structure of the ResNet-50 [He et al., 2015].*

فهي تبدأ بتلافف 7×7 يحول المدخل ذي القنوات الثلاثة لصورة قناتها 64 بنصف الأبعاد متبعةً بأربع أقسام قطعية متجاوزة ومن العجيب ألا تصغيرَ في القسم الأول فقط زيادة في عدد القنوات بالمعامل 5. المخرج في آخر قطعة متجاوزة هو 2048×7×7 والذي يحول لمتجهٍ بعُده 2048 بتجميع توسيطيّ لنواةٍ حجمها 7×7 ثم تعالج في طبقة كامل الاتصال للحصول على المُحُض النهائية وهنا هي 1000 صنف.

كما وسبق الاشارة له ففي عدة تطبيقات تحديدًا في معالجة اللغة الطبيعية تستفيد كثيرًا من النماذج المحتوية على آليات الانتباه. الهيكلية المختارة لهذه المهمات كان لها تأثير عظيم في التقدمات الأخير في التعلم العميق هي هيكلية المحول [ورقة جوجل].

المحوّل

المحول الأصلي المصور تاليًا

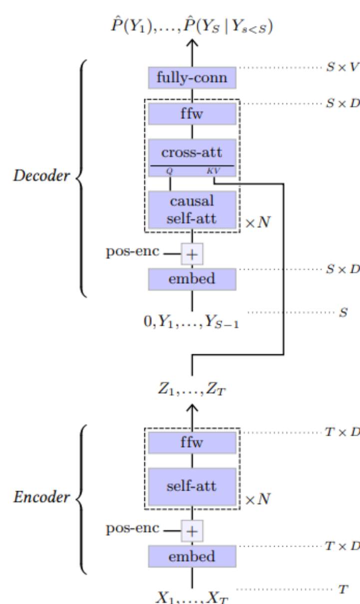
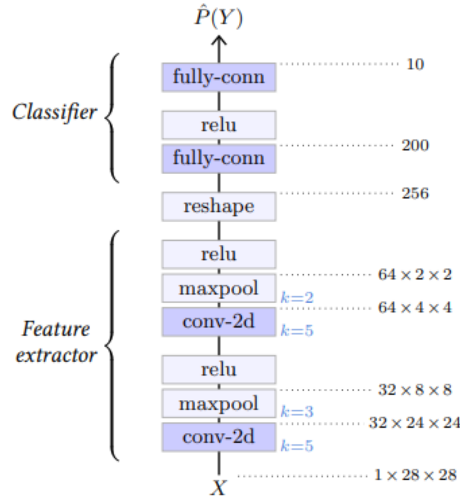


Figure 5.7: *Original encoder-decoder Transformer model for sequence-to-sequence translation [Vaswani et al., 2017].*

مصممة للترجمة مقاطَرة السلاسل (سلسلة لسلسلة). تتضمن (تتركب) من مشفر يعالج السلسلة المُدخلة ليحصل على تمثيل واضح ومفكك نكوصي يولد وحدة لغوية (كليمة) في السلسلة الناتجة مُعطى تمثيل المشفر للمُدخل والمخرجات المولدة للآن.

كما في الشبكات التلافيفية المتجاوزة في



وقد مرت

كلًّا من المشفر والمفكك في المحولات سلاسل مركبة من قطع مبنية باتصالات متجاوزة.

- قطعة التغذية الأمامية المصوّرة أول الصورة

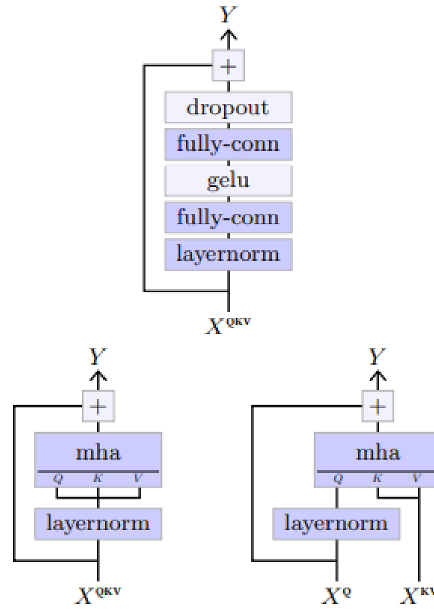


Figure 5.6: Feed-forward block (top), self-attention block (bottom left) and cross-attention block (bottom right). *These specific structures proposed by Radford et al. [2018] differ slightly from the original architecture of Vaswani et al. [2017], in particular by having the layer normalization first in the residual blocks.*

هي طبقة مخفية MLP شعصبونية يتبعها طبقة نظم تحدّث التمثيلات كل موقع منفصلًا.

- قطعة الانتباه الذاتي المصورّة أسفل الصورة السابقة هي طبقة الانتباه عديد الرؤوس والتي تجمّع المعلومات شموليًّا سامحةً لأي موقع أن يجمع معلومات من أي آخر متبوعة بطبقة نظم هذه القطعة يمكن جعلها سببها باستعمال التقنيع في طبقة الانتباه كما وصف في المعادلة السابقة.

- قطعة الانتباه المتقاطع كما هي في الصورة السابقة مشابهة باستثناء أن المدخل سلسلتان واحدة للمستعلمات وواحدة للمفاتيح والقيَم.

المشفّر في المحوّل (في الصورة الأولى) تشفر المدخل من الوحدات المقطّعة $X_1,...X_T$ مع طبقة التضمين وإضافة تشفير الموقع (الموقعة) (انظر السوابق) قبل معالجة بعدة قطع انتباهات ذاتية لتوليد التمثيل المعدّل $Z_1,...,Z_T$.

المفكك يأخذ مدخلًا سلسلة $Y_1,...,Y_{S-1}$ للوحدات الناتجة للآن، كذلك يشابه أنه يشفّرهم عبر طبقة التضمين تضيف تشفيرًا موقعيًّا وتعالجها خلال مسببة الانتباه الذاتي والمتقاطع قطعًا ينتج مُحُضًا متوقعا الوحدات التالية.

هذه القطع بالانتباه المتقاطع تحسب المفاتيح والقيم من نتائج تمثيلات المشفّر $Z_1,...,Z_T$ والتي تسمح السلسلة الناتجة لتكون دالة للسلسلة الأصلية $X_1,...X_T$

كما رأينا سابقا كونه سببي يضمن أن يدرّب باستصغار الأنطروب المتقاطع المجموعة عبر كل السلسلة.

المحوّل المُدَّرب المولّد (متمم)

هذا (المتمم) كما في الصورة

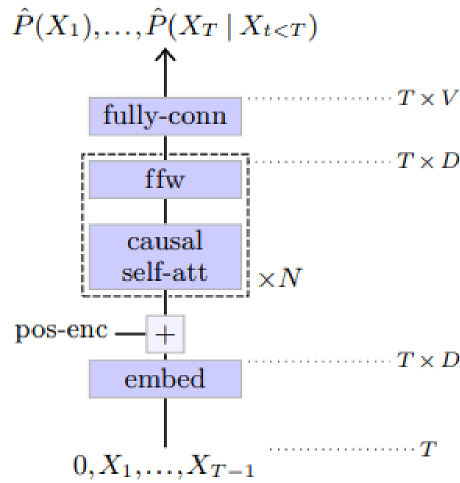


Figure 5.8: *GPT model [Radford et al., 2018].*

هو نموذج نكوصيٌّ بحتٌ يحتوي على طبقات انتباه سببي وعليه فهو سببي من مفكك المحوّل الأصلي.

هذا النوع من النماذج يتحجّم بشكل رائع مبهر واصلًا لمئات بلايين المتغيرات المندرّبة وسنعود له لاحقًا [ان شاء الله]

المحوّل الرؤيوي (البصري، الصوري)

المحولات قد استخدمت لتصنيف الصور في المحول الرؤيوي (المحؤلي) انظر الصورة

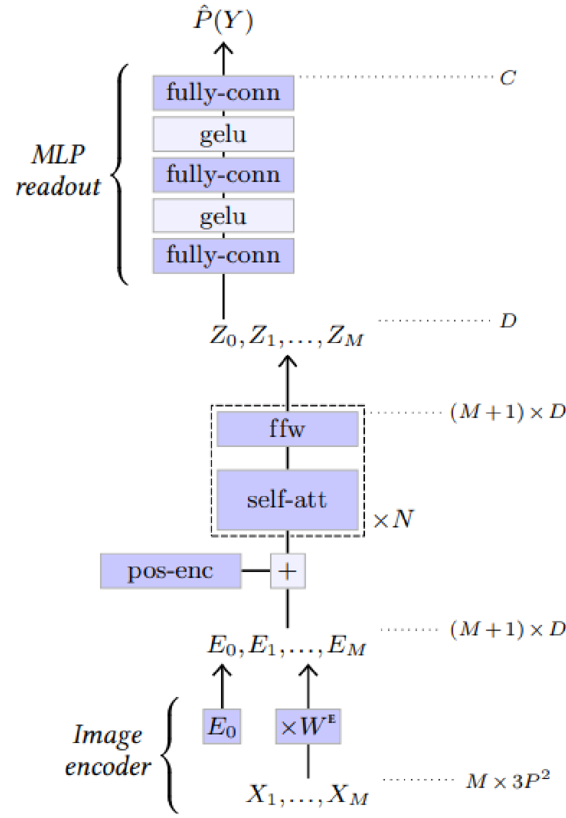


Figure 5.9: *Vision Transformer model [Dosovitskiy et al., 2020].*

هو يقطّع المدخل الصوري ثلاثي القنوات لـ M من القطع بأبعاد P×P والتي تُسطح (تسوّى، تدك، تدح) لتوليد سلاسل متجهات $X_1, \ldots, X_M$ جمها $M \times 3P^2$ هذه السلسلة تُضرب بمصفوفة مندربة W^E جمها $M \times 3P^2$ يُطَبَّق في سلسلة $M \times D$ والتي تُضَمّ في متجه مندرب E_0. الناتج $(M+1) \times D$ سلسلة له $E_0, \ldots, E_M$ ثم تعالج عبر عدة قطع من الانتباه الذاتي انظر السابق.

العُنصر الأول Z_0 في السلسلة الناتجة تؤشر لـ E_0 وليست مرتبطة مع أي قطعة في الصورة أخيرًا تعالجة بطبقتين شعصبونات للحصول على المحض النهائية C ووحدة كهذه (صورية) تضاف لتوقع الصنف (النوع) المتوقّع في نموذج المتفهّم ويشار له باسم الوحدة المبتدأة CLS.

القسم الثالث: التطبيقات

الفصل السادس: التوقّع

[تنبيه [29]]

الصنف الأول من التطبيقات مثل التعرف الوجهي والتحليل المشاعري وتحديد الأشياء والتعرف الصوتي تتطلب توقّع قيمة مجهولة من إشارة متوفّرة.

إزالة تشويش الصورة

تطبيق مباشر للنماذج العميقة في معالجة الصور هو **استعادة الصورة من حالات التدهور** عبر الاستفادة من التكرارية في البنية الإحصائية للصور. فمثلًا، يمكن تلوين بتلات زهرة دوّار الشمس في صورة رمادية بدرجة عالية من الثقة، كما يمكن تصحيح نسيج شكل هندسي مثل الطاولة في صورة منخفضة الإضاءة ومليئة بالضوضاء من **خلال إجراء متوسط على مساحة واسعة يُرجَّح أن تكون متجانسة.**

المُرمِّز التلقائي لإزالة الضوضاء (Denoising Autoencoder) هو نموذج يأخذ إشارة متدهورة X~ كمدخل، ويُنتج تقديرًا للإشارة الأصلية X.
في حالة الصور، يكون عادةً شبكة التفافية (Convolutional Network) قد تتضمن **وصلات تخطّي (متجاوزة) (Skip-Connections)**، خصوصًا لدمج التمثيلات على نفس الدقة التي يتم الحصول عليها في المراحل المبكرة والمتأخرة من النموذج، بالإضافة إلى **طبقات الانتباه (Attention Layers)** لتسهيل التعامل مع العناصر البعيدة عن بعضها البعض.

يتم تدريب مثل هذا النموذج عبر جمع عدد كبير من العينات النظيفة مع أقرانها من المدخلات المتدهورة. يمكن الحصول على هذه الأخيرة في ظروف متدهورة مثل الإضاءة المنخفضة أو ضعف التركيز، أو إنشاؤها خوارزميًا، مثل تحويل العينة النظيفة إلى صورة رمادية، أو تقليل حجمها، أو ضغطها بشدة باستخدام طريقة ضغط ضائعة (Lossy Compression).

[29] من هنا إلى آخر الكتاب تم استعمال الترجمة الآلية لإنهائه نظراً لضيق الوقت ولأنها مواضيع عامّة مبنية على فهم ما قبلها

إجراء التدريب القياسي للمُرمِّزات التلقائية لإزالة الضوضاء يستخدم **خسارة متوسط مربعات الخطأ (MSE Loss)** على جميع البكسلات، حيث يسعى النموذج إلى حساب **أفضل صورة نظيفة متوقعة بالمتوسط** عند إعطائه الصورة المتدهورة، أي: القيمة المتوقعة لـ X عندما تكون X~ معطاة.

لكن هذه القيمة قد تكون إشكالية عندما لا تكون X محددة بالكامل بواسطة X~، إذ قد ينتج عن ذلك أجزاء من الإشارة المولدة تبدو غير واقعية أو ضبابية باعتبارها مجرد متوسط تقريبي. $\mathbb{E}[X \mid \tilde{X}]$.

تصنيف الصور

تصنيف الصور (Image Classification) هو أبسط استراتيجية لاستخلاص الدلالات (Semantics) من الصورة، ويقوم على التنبؤ بفئة واحدة من بين عدد محدود ومُعرَّف مسبقًا من الفئات، وذلك اعتمادًا على صورة مدخلة.

النماذج القياسية لهذه المهمة هي **الشبكات الالتفافية (Convolutional Networks)** مثل **ResNets**، والنماذج المعتمدة على **الانتباه (Attention-based Models)** مثل **ViT** . هذه النماذج تُنتج متجهًا من القيم المحضة (Logits) بعدد أبعاد يساوي عدد الفئات.

إجراء التدريب هنا ببساطة يقوم على **تصغير خسارة الانتروبيا المتقاطعة (Cross-Entropy Loss)**. وعادةً يمكن تحسين الأداء باستخدام **زيادة البيانات (Data Augmentation)**، والتي تتكون من تعديل عينات التدريب عبر تحولات عشوائية مصمَّمة يدويًا لا تُغيّر المحتوى الدلالي للصورة، مثل: القص (Cropping)، تغيير الحجم (Scaling)، العكس أو المرآة (Mirroring)، أو تغييرات الألوان (Color Changes).

مهمة أكثر تعقيدًا لفهم الصور هي **اكتشاف الكائنات**، حيث يكون الهدف هو التنبؤ بفئات الكائنات ومواضعها في صورة مدخلة.

يُصاغ موضع الكائن من خلال **أربع إحداثيات (x1, y1, x2, y2)** تمثل المستطيل المحيط بالكائن (Bounding Box). أما القيمة الحقيقية (Ground Truth) المرتبطة بكل صورة تدريب فهي عبارة عن **قائمة بهذه المربعات المحيطة**، كل واحد منها مُعنون -موسوم- (Labeled) بالفئة التي ينتمي إليها الكائن الموجود بداخله.

المنهج القياسي لحل هذه المهمة — مثلًا باستخدام **كاشف المثال الواحد ((Single Shot Detector, SSD** [[Liu et al., 2015 — هو استعمال شبكة عصبية التفافية (Convolutional Neural Network) تنتج سلسلة من تمثيلات الصورة **Zs** ذات أبعاد **Ds** × **Hs** × **Ws** حيث s = 1,…,S، مع انخفاض تدريجي في الدقة المكانية Hs × Ws حتى تصل إلى 1×1 عندما يكون s = S (انظر الشكل التالي).

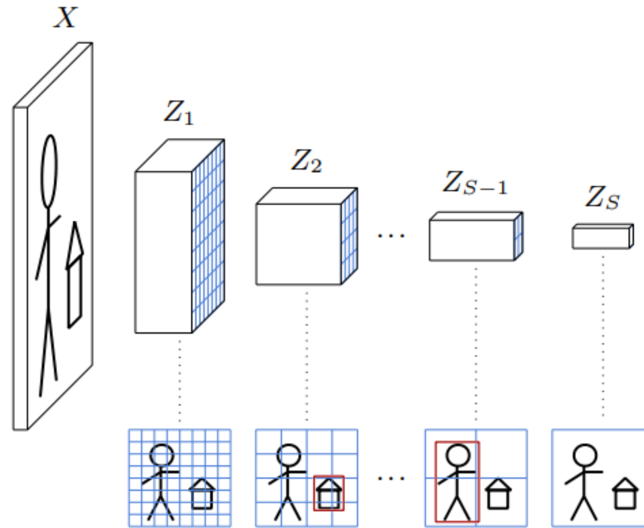


Figure 6.1: *A convolutional object detector processes the input image to generate a sequence of representations of decreasing resolutions. It computes for every* h, w, *at every scale* s, *a pre-defined number of bounding boxes whose centers are in the image area corresponding to that cell, and whose sizes are such that they fit in its receptive field. Each prediction takes the form of the estimates* $(\hat{x}_1, \hat{x}_2, \hat{y}_1, \hat{y}_2)$, *represented by the red boxes above, and a vector of* $C+1$ *logits for the* C *classes of interest, and an additional "no object" class.*

كل واحد من هذه المرصوصات يغطي الصورة المدخلة بالكامل، بحيث أن الفهارس (h, w) تقابل تقسيم شبكة الصورة إلى مربعات منتظمة تصبح أكبر حجمًا كلما زاد s.

كما هو موضح في القسم السابق، وبسبب تعاقب الطبقات الالتفافية، فإن متجه الميزة $(Z_s[0,h,w],\ldots,Z_s[D_s-1,h,w])$ هو واصف (Descriptor) لمنطقة من الصورة تُسمى **المجال الاستقبالي (Receptive Field)**. هذا المجال يكون أكبر من المربع الذي يقابله لكنه يتمركز عليه. والنتيجة هي إمكانية ربط أي مربع محيط (Bounding Box) بالإحداثيات (x1, y1, x2, y2) بشكل غير ملتبس مع قيمة محددة من (s, h, w)، والتي يتم تحديدها على التوالي بواسطة:

- أكبر بُعد بين (x2 − x1, y2 − y1)،
- المركز العمودي ((y1 + y2) ÷ 2)،
- المركز الأفقي ((x1 + x2) ÷ 2).

يتم إنجاز عملية **الاكتشاف/التحديد (Detection)** عبر إضافة **S طبقات التفاف إضافية (Convolutional Layers)**، بحيث تقوم كل طبقة بمعالجة تمثيل معين Zs، ثم حساب **إحداثيات المربع المحيط (Bounding Box)** والقيم المرتبطة به (Logits) لكل موضع (h, w) في التنسور.

إذا كان لدينا **C فئة من الكائنات**، فسيكون هناك **C + 1 قيمة Logit**، حيث تمثل الإضافة الأخيرة فئة "لا يوجد كائن" (No Object). وبالتالي، فإن كل طبقة التفاف إضافية تنتج **4 + 1 + C قناة مخرجات** (4 لإحداثيات المربع + C للفئات + 1 لفئة اللا شيء).
خوارزمية **SSD** بشكل خاص تُولّد **عدة مربعات محيطة** لكل موضع (s, h, w)، حيث يُخصص كل منها لنطاق محدد مسبقًا من **نسب الأبعاد (Aspect Ratios)**.

إعداد مجموعات بيانات التدريب الخاصة باكتشاف الكائنات مكلف، لأن عملية وضع المربعات المحيطة تتطلب تدخلاً بشريًا بطيئًا. وللتخفيف من ذلك، يُستخدم الأسلوب القياسي القائم على **إعادة الضبط (Fine-Tuning)** لنموذج التفاف تم تدريبه مسبقًا على مجموعة بيانات ضخمة للتصنيف مثل **VGG-16** (كما في النسخة الأصلية من SSD)، ثم استبدال طبقاته الكاملة التوصيل (Fully-Connected) بطبقات التفاف إضافية. والمثير للدهشة

(والمُدهش) أن النماذج المدربة فقط على مهمة التصنيف تتعلم تمثيلات للميزات يمكن إعادة استخدامها في مهمة اكتشاف الكائنات، رغم أن هذه المهمة تتضمن أيضًا انحدارًا (نكوصًا) كميًا **(Regression)** لإحداثيات هندسية.

أثناء التدريب، يتم ربط كل مربع محيط حقيقي (Ground Truth Bounding Box) بالقيمة (s, h, w) الخاصة به، وينتج عن ذلك عنصر في الخسارة يتكون من:

- **خسارة انتروبيا متقاطعة (Cross-Entropy Loss)** للقيم المحضَة Logits
- و **خسارة انحدار (مثل MSE)** لإحداثيات المربع.

أما أي موضع (s, h, w) لا يُطابق مربعًا محيطًا حقيقيًا، فإنه يساهم فقط بخسارة انتروبيا متقاطعة تهدف إلى التنبؤ بالفئة "لا يوجد كائن" (No Object).

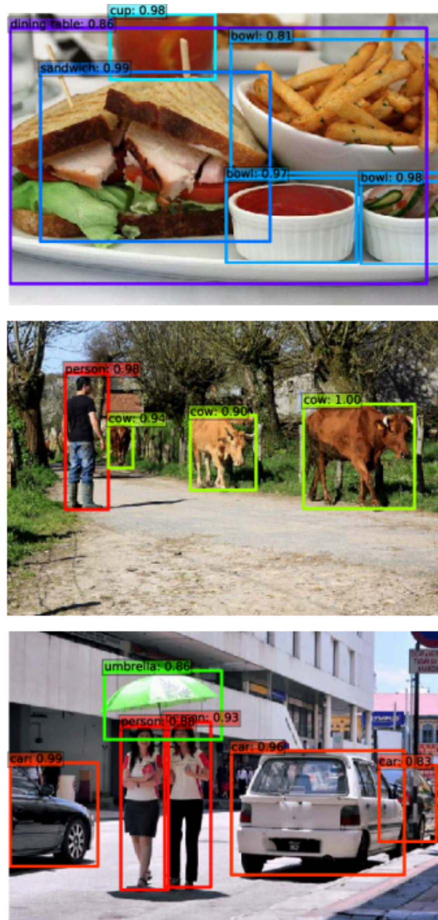


Figure 6.2: *Examples of object detection with the Single-Shot Detector [Liu et al., 2015].*

التجزئة الدلالية (Semantic Segmentation)

أدق مهمة للتنبؤ في فهم الصور هي **التجزئة الدلالية**، والتي تقوم على التنبؤ بفئة الكائن التي ينتمي إليها كل بكسل في الصورة. يتم ذلك باستخدام شبكة عصبية التفافية قياسية تنتج خريطة التفاف (Convolutional Map) بعدد قنوات يساوي عدد الفئات، بحيث تحمل كل قناة القيم التقديرية (Logits) الخاصة بكل بكسل.

على الرغم من أن شبكة قياسية مثل **ResNet** يمكن أن تُولّد مخرجات كثيفة بنفس دقة الصورة المدخلة، إلا أن هذه المهمة — مثل اكتشاف الكائنات — تتطلب العمل على **مقاييس متعددة (Multi-Scales)**. السبب هو أنه يجب ضمان أن أي كائن (أو جزء فرعي مهم منه)، بغض النظر عن حجمه، يتم تمثيله في مكان ما داخل النموذج عبر متجه ميزات في موضع تنسور واحد.

لهذا السبب، تعتمد البنى القياسية لهذه المهمة على **تصغير تدريجي للصورة (Downscaling)** عبر سلسلة من طبقات الالتفاف، وذلك لتوسيع **المجال الاستقبالي (Receptive Field)** للتنشيطات، ثم **إعادة تكبيرها (Upscaling)** عبر سلسلة من طبقات الالتفاف المنقولة (Transposed Convolutions) أو عبر طرق أخرى مثل **الاستيفاء الخطي الثنائي (Bilinear Interpolation)**، من أجل إنتاج تنبؤات عالية الدقة.

لكن، بنية "تصغير–تكبير" صارمة لا تسمح بالتنبؤ بدقة عالية جدًا، لأن كل الإشارة تمرّ في مرحلة ما عبر تمثيل منخفض الدقة. لحل هذه المشكلة:

- هناك نماذج تطبق **وصلات تخطّي (Skip Connections)** من الطبقات عند دقة معينة (قبل التصغير) إلى طبقات عند نفس الدقة (بعد التكبير)، مثل [Long et al., 2014; Ronneberger et al., 2015].
- وهناك نماذج أخرى تطبق ذلك **بالتوازي** بعد العمود الفقري الالتفافي (Convolutional Backbone)، حيث يتم دمج التمثيلات الناتجة من مقاييس متعددة بعد إعادة التكبير، قبل إجراء التنبؤ النهائي لكل بكسل [Zhao et al., 2016].

يتم التدريب باستخدام **خسارة الانتروبيا المتقاطعة (Cross-Entropy Loss)** مجموعًا عبر جميع البكسلات. وكما في مهمة اكتشاف الكائنات، يمكن بدء التدريب من شبكة مدربة مسبقًا على مهمة

تصنيف صور ضخمة، للتعويض عن محدودية توفر بيانات الحقيقة الأرضية (Ground Truth) الخاصة بالتجزئة.



Figure 6.3: *Semantic segmentation results with the Pyramid Scene Parsing Network [Zhao et al., 2016].*

التعرّف على الكلام هو عملية تحويل عيّنة صوتية إلى تسلسل من الكلمات. تاريخيًا، وُجدت العديد من الأساليب لمعالجة هذه المشكلة، لكن أحد الأساليب الحديثة والمبسّطة مفاهيميًا هو ما قدّمه **Radford وآخرون (2022)**، حيث تم صياغة المهمة على أنها **ترجمة تسلسل-إلى-تسلسل (Sequence-to-Sequence Translation)** ثم حلّها باستخدام **محول قائم على الانتباه (Attention-based Transformer)** كما هو موضح في القسم 5.3.

في هذا النموذج، يتم أولًا تحويل الإشارة الصوتية إلى **مطيافية (Spectrogram)**، وهي سلسلة أحادية البعد بحجم T × D، بحيث يمثّل كل موضع زمني متجهًا من الطاقات عبر D نطاقات ترددية. أما النص المرتبط بالصوت فيُشفَّر باستخدام **مجزّئ BPE** (المزبّت [المزوج الباتِيّ[30]]) (انظر القسم 3.2).

تُعالج المطيافية عبر عدد من طبقات الالتفاف أحادية البعد (1D Convolutions)، ثم يُمرَّر التمثيل الناتج إلى **المُشفّر (Encoder)** في المحول. بعد ذلك، يقوم **المُولّد (المفكك) (Decoder)** مباشرة بإنتاج تسلسل متقطّع من الرموز (Tokens)، كل منها يقابل إحدى المهام المدروسة أثناء التدريب.
المهام متعددة وتشمل:

- نسخ الكلام إلى نص (بالإنجليزية أو بغيرها).
- الترجمة من أي لغة إلى الإنجليزية.
- اكتشاف التسلسلات غير الكلامية، مثل الموسيقى الخلفية أو الضوضاء المحيطة.

هذا النهج يتيح الاستفادة من **مجموعات بيانات ضخمة جدًا** تجمع أنواعًا مختلفة من المصادر الصوتية مع حقائق أرضية متنوعة.

من الجدير بالذكر أنه بالرغم من أن الهدف النهائي لهذا النهج هو إنتاج ترجمة حتمية قدر الإمكان استنادًا إلى الإشارة الصوتية، فإنه من الناحية الشكلية عبارة عن **توليد نص موزّع احتماليًا مشروطًا بعينة صوتية**، وبالتالي يُعد عملية تركيب (Synthesis). في الواقع، فإن المُولّد (Decoder) هنا يشبه إلى حد كبير **النموذج التوليدي** الموضح في القسم السابق

[30] الرقم الثنائي binary قد ينحت له رقث مثل bit أما byte فأصلها أنها لنفس الرقم ثم استعملت في الثماني فممكن نحته رقثم بإضافة الميم

تمثيلات النص-الصورة (Text-Image Representations)

أحد الأساليب القوية لفهم الصور هو **تعلّم تمثيلات متسقة لكل من الصور والنصوص**، بحيث يتم تمثيل الصورة أو وصفها النصي بنفس متجه الخصائص (Feature Vector).

من أشهر النماذج في هذا المجال نموذج **CLIP (Contrastive Language-Image Pre-training)** الذي اقترحه **Radford** وآخرون [2021]، حيث يجمع بين:

- **مُشفّر صور (Image Encoder) f** وهو شبكة **ViT**.
- **مُشفّر نصوص (Text Encoder) g** وهو شبكة **GPT** (انظر القسم 5.3 لكليهما).

لإعادة استخدام GPT كمُشفّر نصوص بدلاً من كونه نموذجًا توليديًا تسلسليًا (Autoregressive)، أضاف الباحثون رمز "نهاية الجملة" إلى تسلسل الإدخال، وتم استخدام تمثيل هذا الرمز في الطبقة الأخيرة كـ **تضمين (Embedding)**. أبعاده تكون عادة بين 512 و 1024 اعتمادًا على التهيئة.

تم تدريب هذين النموذجين معًا من الصفر باستخدام **400 مليون زوج صورة–نص (ik, tk)** جُمعت من الإنترنت. ويُتبع في التدريب أسلوب الانحدار العشوائي المصغّر (Mini-Batch SGD)، لكن مع **خسارة تقابلية (Contrastive Loss)**:

- يتم حساب التضمينات لكل صورة ولكل نص في الدفعة (Batch) ذات الحجم N.
- ثم تُحسب **تشابهات جيب التمام (Cosine Similarity)** ليس فقط بين النص والصورة في كل زوج، بل أيضًا عبر جميع الأزواج.
- هذا يُنتج مصفوفة N × N من درجات التشابه: lm,n = f(im) · g(tn).
- يُدرّب النموذج باستخدام **خسارة الانتروبيا المتقاطعة (Cross-Entropy Loss)** بحيث تكون القيمة ln,n (التشابه بين الصورة والنص الصحيح) أكبر بوضوح من أي تشابه آخر ln,m أو lm,n عندما m ≠ n.

بعد التدريب، يمكن استخدام هذا النموذج في **التصنيف بدون أمثلة تدريبية (Zero-Shot Prediction)**: أي تصنيف صورة جديدة من خلال تعريف فئات مرشحة على شكل أوصاف نصية، ثم حساب تشابه التضمين الخاص بالصورة مع تضمين كل وصف نصي (انظر الشكل 6.4).

ميزة إضافية مهمة: بما أن الأوصاف النصية غالباً ما تكون مفصّلة، فإن النموذج يضطر إلى تعلم **تمثيلات أكثر ثراءً للصور** تلتقط إشارات تتجاوز ما يكفي للتصنيف التقليدي. هذا ينعكس في أداء ممتاز على مجموعات بيانات صعبة مثل **ImageNet Adversarial [Hendrycks et al., 2019]**، والتي صُمّمت خصيصًا لتشويش أو إزالة الإشارات التي تعتمد عليها النماذج القياسية.

يُعَدّ أحد الأساليب القوية لفهم الصور هو **تعلّم تمثيلات مشتركة ومتّسقة للنص والصورة**، بحيث يمكن تمثيل الصورة أو الوصف النصي لها **بنفس متجه السمات.**

يقوم نموذج **التدريب المتباين للّغة والصورة (CLIP)** الذي اقترحه **رادفورد وآخرون (2021)** بدمج **مُرمِّز صور** يُرمز له بـ **f** (وهو من نوع **ViT – Vision Transformer**) مع **مُرمِّز نص** يُرمز له بـ **g** (وهو من نوع **GPT – Generative Pre-trained Transformer**). (انظر القسم السابق لشرح كلا النموذجين).

ولتحويل نموذج GPT من كونه نموذجًا **توليديًا تسلسليًا** إلى **مُرمِّز نصّي**، أضاف الباحثون **رمزًا خاصًا يدل على "نهاية الجملة"** إلى تسلسل الإدخال، ثم استخدموا **تمثيل هذا الرمز في الطبقة الأخيرة** كـ **تضمين (embedding)** للجملة.
تتراوح **أبعاد هذا التضمين** بين **512 و1024**، وفقًا لإعداد النموذج.

تمّ تدريب هذين النموذجين **من الصفر** باستخدام **مجموعة بيانات ضخمة تضم 400 مليون زوج من الصور والنصوص (ik, tk)** جُمعت من الإنترنت.
ويتم التدريب باستخدام **الانحدار العشوائي المصغّر (mini-batch SGD)** مع **دالّة خسارة متباينة (contrastive loss)**.

يُحسب التضمين لكل صورة ولكل نص من الأزواج **N** في الدفعة التدريبية، ثم يُحسب **تشابه جيبي (cosine similarity)** ليس فقط بين كل زوج (صورة، نص) متطابق، بل أيضًا **بين جميع الأزواج المختلفة**، مما ينتج **مصفوفة تشابه بحجم N×N**:

$$l_{m,n} = f(i_m) \cdot g(t_n), m = 1, \ldots, N, n = 1, \ldots, N.$$

ثم يُدرّب النموذج باستخدام **خسارة الانتروبيا التقاطعية (cross-entropy)** بحيث تكون —لكل n— القيم (l1,n … lN,n) ممثلة كدرجات logits تتنبأ بالزوج الصحيح n، وكذلك (ln,1 … ln,N).
ويُفترض أن **التشابه بين الصورة والنص المطابق (ln,n)** يجب أن يكون **أكبر بوضوح** من التشابهات **ln,m** و**lm,n** عندما **n ≠ m**.

بعد التدريب، يمكن استخدام النموذج لتنفيذ ما يُعرف بـ التنبؤ دون تدريب مسبق **(zero-shot prediction)**، أي تصنيف الإشارات أو الصور دون وجود أمثلة تدريبية، وذلك عبر تعريف **مجموعة من الفئات المحتملة** بوصف نصي لكل فئة، ثم حساب تشابه التضمين بين الصورة وتمثيل **كل وصف نصي** (انظر الشكل 6.4).

وبما أنّ الأوصاف النصية غالبًا ما تكون **تفصيلية وغنية بالمعاني**، فإنّ مثل هذا النموذج يضطر إلى **التقاط تمثيلات أعمق وأغزر** للصور، متجاوزًا ما تحتاجه مهام مثل التصنيف فقط.
وهذا ما يجعله يُظهر أداءً ممتازًا في مجموعات بيانات صعبة مثل **ImageNet Adversarial** (هندريكس وآخرون، **2019**)، والتي صُمّمت خصيصًا لتقليل أو إزالة الإشارات التي تعتمد عليها النماذج التقليدية.

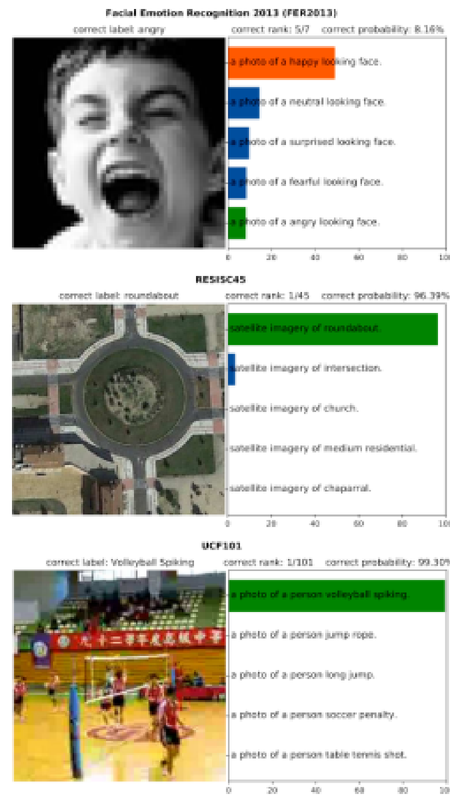


Figure 6.4: *The CLIP text-image embedding [Radford et al., 2021] allows for zero-shot prediction by predicting which class description embedding is the most consistent with the image embedding.*

التعلّم بالتعزيز (Reinforcement Learning)

تُعالج كثيرٌ من المسائل — مثل ألعاب الاستراتيجية أو التحكم بالروبوتات — من خلال ما يُعرف بـ **عملية الحالة المتقطعة زمنياً (discrete-time state process)**، والمُقترنة **بعملية مكافأة (reward process)** يمكن **التحكم بها عبر اختيار أفعال (Actions)** في كل خطوة زمنية.

إذا كانت الحالة **St ماركوفية (Markovian)** — أي أنّها تحتوي وحدها على جميع المعلومات اللازمة للتنبؤ بالمستقبل دون الحاجة إلى معرفة الحالات السابقة — فإن النظام يُسمّى **عملية اتخاذ قرار ماركوفية (Markov Decision Process, MDP)**.

في إطار الـ **MDP**، الهدف هو إيجاد **سياسة (policy)** يرمز لها بـ **π** بحيث يكون الفعل في كل لحظة **At = π(St)** مصمماً **لتحقيق أعلى قيمة متوقعة للعائد الكلي (expected return)**، والذي يُحسب عادةً على شكل مجموع المكافآت المخفّضة عبر الزمن:

$$\mathbb{E}\left[\sum_{t\geq 0}\gamma^t R_t\right],$$

for a discount factor $0 < \gamma < 1$.

حيث γ هو **عامل الخصم (discount factor)** بين 0 و1، ويُستخدم لتقليل وزن المكافآت البعيدة زمنياً.

هذا هو الإطار الكلاسيكي لـ **التعلّم بالتعزيز (Reinforcement Learning)**.
ولتطبيقه عملياً، يُعرّف ما يُعرف بـ **دالّة القيمة المثلى للحالة والفعل (optimal state-action value function)**، والمُرمّز لها بـ **Q(s, a)**، وهي **العائد المتوقع عند تنفيذ الفعل a في الحالة s** ثم اتباع السياسة المثلى بعد ذلك.

يمكن عندئذٍ **استخلاص السياسة المثلى** كما يلي:

$$\pi(s) = \text{argmax}_a\, Q(s, a)$$

وتحقق معادلة بيلمان **(Bellman Equation)** تحت افتراض ماركوفية النظام:

$$Q(s,a) = \mathbb{E}\left[R_t + \gamma \max_{a'} Q(S_{t+1},a') \middle| S_t = s, A_t = a\right], \quad (6.1)$$

انطلاقاً من هذه المعادلة، يمكن تصميم إجراء تدريجي لنموذج بارامتري **w ;·,·)Q)**.

تطبيق عملي: ألعاب الفيديو الكلاسيكية (Atari)

لتطبيق هذا الإطار على **ألعاب الفيديو من نوع Atari**، استخدم **Mnih وآخرون (2015)** الحالة **St** كـ **سلسلة من أربع لقطات متتالية (الإطار الحالي وثلاثة سابقة)**، لضمان تقريب خاصية ماركوف.
واستخدموا نموذجًا يُعرف باسم **شبكة القيم العميقة (Deep Q-Network, DQN)**، وهو يتكوّن من **طبقتين التفافيتين (convolutional)** تليهما **طبقة كاملة الاتصال (fully connected)**، بحيث ينتج النموذج **قيمة Q لكل فعل ممكن** — وهي بنية مشابهة لشبكة **LeNet** (انظر 5.2§).

يتم التدريب عبر **اللعب وجمع الحلقات (episodes)** وتخزينها، ثم بناء **دفعات مصغّرة (mini-batches)** من العينات **(sn, an, rn, s′n)** تماثل (St, At, Rt, St+1)، ومن ثم تقليل دالة الخسارة:

$$\mathscr{L}(w) = \frac{1}{N}\sum_{n=1}^{N}\left(Q\left(s_n, a_n; w\right) - y_n\right)^2 \quad (6.2)$$

حيث:

- **yn = rn** إذا كانت الحلقة قد انتهت،
- **(yn = rn + γ * maxₐ Q(s′n, a; w̄** في غير ذلك.

النموذج **w̄** هو **نسخة ثابتة من w**، أي أن **الاشتقاق (gradient)** لا ينتقل خلالها، وذلك لأن **قيمة الهدف yn تمثل التوقع** في معادلة بيلمان، ويُستحسن تثبيت **w̄** لتحسين استقرار التقريب أثناء التدريب.

سياسة الاستكشاف (Exploration Policy)

تُعدّ السياسة المستخدمة لجمع البيانات أمرًا حاسمًا.
اعتمد Mnih وآخرون (2015) على استراتيجية ε-greedy، حيث:

- يُختار فعل عشوائي باحتمال ε (للاستكشاف)،
- ويُختار الفعل الأفضل ($argmax_a$ Q(s, a)) باحتمال 1 − ε (للاستغلال).

إضافة هذه العشوائية ضرورية لتحقيق توازن بين الاستكشاف (exploration) والاستغلال (exploitation).

النتائج

تمّ تدريب النموذج على **عشرة ملايين إطار**، أي ما يُقارب **ثمانية أيام من اللعب المتواصل**.
وبعد التدريب، تمكنت الشبكة من **تقدير قيم الحالات بدقة عالية** (انظر الشكل 6.5)، وحققت **أداءً مماثلًا لأداء البشر في أغلب الألعاب الـ49** التي استخدمت للتحقق التجريبي.

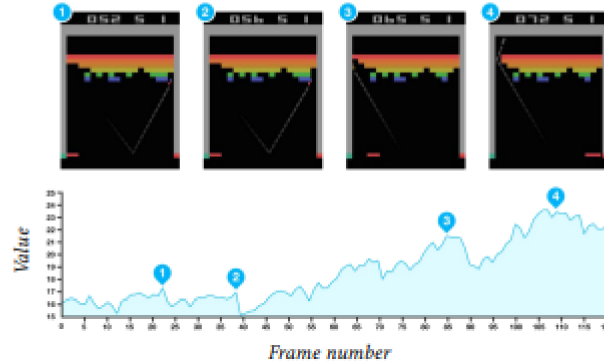


Figure 6.5: *This graph shows the evolution of the state value* $V(S_t) = \max_a Q(S_t, a)$ *during a game of Breakout. The spikes at time points (1) and (2) correspond to clearing a brick, at time point (3) it is about to break through to the top line, and at (4) it does, which ensures a high future reward [Mnih et al., 2015].*

الفصل السابع: التوليد (Synthesis)

الفئة الثانية من تطبيقات النماذج – والمتميّزة عن فئة **التنبؤ (prediction)** – هي **التوليد (synthesis).** ويُقصد بها بناء **نموذج احتمالي (density model)** يتعلّم من **العينات التدريبية** توزيعها الإحصائي، بحيث يتمكن لاحقًا من **توليد عينات جديدة** تشبه تلك التي تدرّب عليها.

توليد النصوص (Text Generation)

المنهج القياسي لتوليد النصوص يعتمد على **نموذج توليدي ذاتي (autoregressive)** قائم على **آلية الانتباه (attention-based model)**.
ومن أبرز النماذج الناجحة في هذا المجال نموذج **GPT** الذي قدّمه **رادفورد وزملاؤه (Radford et al., 2018)**، والذي سبق أن تناولناه في القسم §5.3.

تُستخدم هذه البنية في إنشاء **نماذج ضخمة جدًا**، مثل **GPT-3** من شركة **OpenAI**، الذي يضم **175 مليار معامل** [Brown et al., 2020].
ويتكوّن هذا النموذج من **96 لبنة انتباه ذاتي (self-attention blocks)**، تحتوي كل منها على **96 رأس انتباه (attention heads)**.
ويُعالج النموذج الرموز (tokens) بأبعاد قدرها **12,288**، مع طبقات MLP ذات بُعد خفي **49,512** داخل كل كتلة انتباه.

عند تدريب مثل هذا النموذج على **مجموعة بيانات ضخمة جدًا**، نحصل على ما يُعرف بـ **النموذج اللغوي الضخم (النلض) (Large Language Model – LLM)**، الذي يُظهر **قدرات استثنائية للغاية**.
فإضافةً إلى تعلمه **البنية النحوية والتركيبية للّغة**، يتعيّن عليه أيضًا استيعاب **معرفة متنوعة جدًا** كي يتمكن من التنبؤ الصحيح بالكلمة التالية في عبارات مثل:

"عاصمة اليابان هي ..."
"عندما يُسخَّن الماء إلى 100 درجة مئوية فإنه يتحوّل إلى ..."
"لأن جروها كان مريضًا، كانت جين ..."

تُنتج هذه القدرات ظاهرةً تُعرف باسم **التعلّم من أمثلة قليلة (few-shot learning)**، حيث يمكن للنموذج التكيّف مع مهمة جديدة استنادًا إلى عدد محدود جدًا من الأمثلة، كما هو موضّح في الشكل التالي:

I: I love apples, O: positive, I: music is my passion, O: positive, I: my job is boring, O: negative, I: frozen pizzas are awesome, O: **positive,**
I: I love apples, O: positive, I: music is my passion, O: positive, I: my job is boring, O: negative, I: frozen pizzas taste like cardboard, O: **negative,**
I: water boils at 100 degrees, O: physics, I: the square root of two is irrational, O: mathematics, I: the set of prime numbers is infinite, O: mathematics, I: gravity is proportional to the mass, O: **physics,**
I: water boils at 100 degrees, O: physics, I: the square root of two is irrational, O: mathematics, I: the set of prime numbers is infinite, O: mathematics, I: squares are rectangles, O: **mathematics,**

Figure 7.1: *Examples of few-shot prediction with a 120 million parameter GPT model from Hugging Face. In each example, the beginning of the sentence was given as a prompt, and the model generated the part in bold.*

والأكثر إدهاشًا هو أن هذه النماذج — عند تزويدها **بمطالبات مصمَّمة بعناية (carefully crafted prompts)** — تُظهر قدرات على **الإجابة عن الأسئلة، وحلّ المشكلات، وإجراء سلاسل من التفكير (chain-of-thought reasoning)**، تبدو بشكلٍ لافت قريبة من **التفكير المنطقي عالي المستوى** [Chowdhery et al., 2022؛ Bubeck et al., 2023].

بسبب هذه القدرات المذهلة، تُسمّى هذه النماذج أحيانًا **النماذج الأساسية (القاعديّة، الأصولية) (Foundation Models)** [Bommasani et al., 2021].

ومع ذلك، **رغم احتوائها على كمٍّ هائل من المعرفة**، فقد تكون هذه النماذج **غير ملائمة تمامًا للتطبيقات العملية**، وخصوصًا عند **التفاعل مع المستخدمين البشريين**. ففي العديد من المواقف، نحتاج إلى **استجابات تتبع أنماط حوارية أقرب إلى تفاعلٍ مفيد مع مساعدٍ ذكي**، وهذا **يختلف إحصائيًا عن البيانات الضخمة** التي تم تدريب النموذج عليها، والتي تتألف من **روايات وموسوعات ومنشورات منتديات ومدونات**.

يُعالج هذا التباين عادةً من خلال **إعادة ضبط دقيقة (معيَرة، ضبطنة) (Fine-tuning)** للنموذج اللغوي (انظر §3.6).

والاستراتيجية السائدة حاليًا هي **التعلّم المعزز من التغذية الراجعة البشرية (Reinforcement Learning from Human Feedback – RLHF)** [Ouyang et al., 2022].

وتقوم هذه المنهجية على **إنشاء مجموعات بيانات صغيرة وموسومة يدويًا** عبر **طلب مساهمة بشرية**، إما بكتابة **إجابات نموذجية** أو **تقييم الإجابات التي يولدها النموذج.**

تُستخدم الإجابات المكتوبة مباشرةً في **إعادة ضبط النموذج (Fine-tuning)**،

أما التقييمات فيُبنى منها **نموذج مكافأة (Reward Model)** يتعلم **تقدير جودة الاستجابة**، ثم يُستخدم هذا النموذج كموجِّه لتحديث النموذج اللغوي عبر **أساليب التعلّم المعزز (Reinforcement Learning)** القياسية، بهدف جعله **ينتج إجابات أقرب إلى التفاعل البشري الطبيعي والمفيد.**

توليد الصور (Image Generation)

تم تطوير العديد من الأساليب العميقة لنمذجة التوزيعات الاحتمالية عالية الأبعاد وأخذ عينات منها.
ومن أقوى هذه الأساليب في **توليد الصور** هو الأسلوب القائم على **عكس عملية الانتشار (Diffusion Process)**، ويُعرف هذا النوع من النماذج — بشكل غير دقيق بعض الشيء — باسم **نماذج الانتشار (Diffusion Models)**.

تقوم الفكرة الأساسية على **تعريف عملية تحليلية تقوم بتخريب العينة الأصلية تدريجيًا**، أي تحويل التوزيع المعقد والمجهول للبيانات إلى **توزيع بسيط ومعروف** (مثل التوزيع الطبيعي).
ثم تُدرَّب شبكة عميقة على **عكس هذا التخريب** خطوةً بخطوة لاستعادة الصورة الأصلية [Ho et al., 2020].

يُحدَّد عدد الخطوات **T** مسبقًا، وتُعرّف عملية الانتشار على شكل توزيع احتمالي عبر سلسلة من الصور **x_0, x_1, ..., x_T** كما يلي:
يُختار **x_0** عشوائيًا من مجموعة البيانات، ثم تُولَّد كل صورة لاحقة وفق توزيع مشروط **p(x_{t+1} | x_t)** معرفٍ تحليليًا بحيث **يزيل تدريجيًا البنية الأصلية للصورة x_0**.
يُفترض أن يصل هذا التخريب في النهاية إلى حدّ يجعل التوزيع **p(x_T)** بسيطًا ومعروفًا (عادةً توزيعًا طبيعيًا قياسيًا يمكن أخذ عينات منه مباشرة).

على سبيل المثال، في نموذج **Ho et al. (2020)**، تُطبع البيانات لتصبح بمتوسط 0 وتباين 1، ثم تُضاف **ضوضاء بيضاء صغيرة** عند كل خطوة مع إعادة تطبيع التباين إلى 1.
يؤدي ذلك إلى تقليص تأثير الصورة الأصلية **x_0** بشكلٍ أسيّ، بحيث يمكن تقريب توزيع **x_t** لاحقًا بتوزيعٍ طبيعي.

يُستخدم **نموذج التنقية (Denoiser)** — وهو شبكة عميقة — لتقريب التوزيع العكسي:
(f(x_{t−1}, x_t, t; w) ≈ p(x_{t−1} | x_t).
ويُظهر تحليل تقريبي باستخدام **الحد الأدنى التبايني (variational bound)** أنَّه إذا كانت هذه العملية العكسية دقيقة بما فيه الكفاية، فإن أخذ عينة **(x_T ~ p(x_T)** ثم تطبيق **T** خطوة من التنقية عبر **f** يؤدي إلى الحصول على **x_0** تتبع توزيع البيانات الأصلية **(p(x_0)**.

يُدرَّب النموذج **f** عبر توليد عدد كبير من سلاسل العينات **$(x_0^{(n)}, \ldots, x_T^{(n)})$**، واختيار خطوة **$t$** في كل سلسلة، ثم **تعظيم مجموع احتمالات $\log f(x_{t-1}^{(n)}, x_t^{(n)}, t; w)$**.

في حالة **(Ho et al. 2020)**، يكون شكل التوزيع العكسي على النحو التالي:

$$x_{t-1} \mid x_t \sim \mathcal{N}(x_t + f(x_t, t; w), \sigma_t) \qquad \text{(المعادلة 7.1)}$$

حيث **σ_t** محددة تحليليًا.

عمليًا، يبدأ هذا النوع من النماذج بـ **ضوضاء عشوائية خالصة**، ثم **"يتخيّل" تدريجيًا البنى العامة** داخل هذه الضوضاء، قبل أن **يعزّز التفاصيل خطوةً بخطوة** حتى تظهر صورة واقعية متكاملة.

ويمكن **توسيع هذه المقاربة لتوليد الصور المشروطة بالنصوص** (Text-Conditioned Image Generation)، أي لتوليد صور تتوافق مع وصف لغوي.
فمثلًا، في عمل **(Nichol et al. 2021)**، أُضيف إلى **متوسط التوزيع العكسي** في المعادلة (7.1) **انحياز (bias)** يُوجّه عملية التوليد نحو **زيادة تطابق الصورة الناتجة مع النص الوصفي** وفقًا لمقياس **CLIP** (انظر §6.6).

وبذلك، تجمع نماذج الانتشار الحديثة بين **القوة الإحصائية في النمذجة والتحكم الدلالي عبر اللغة**، مما يجعلها من **أقوى الأساليب الحالية لتوليد الصور عالية الجودة والمطابقة للوصف النصي.**

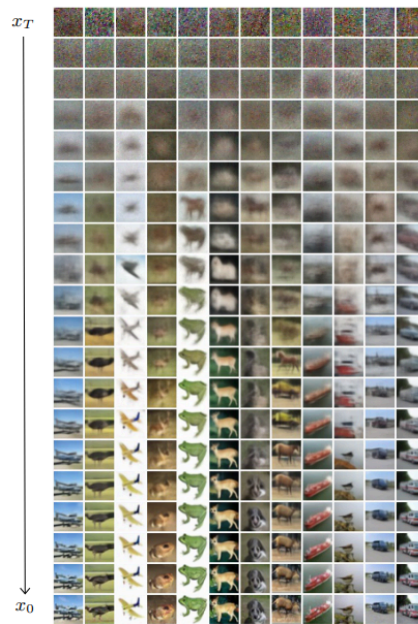


Figure 7.2: *Image synthesis with denoising diffusion [Ho et al., 2020]. Each sample starts as a white noise x_T (top), and is gradually de-noised by sampling iteratively $x_{t-1} \mid x_t \sim \mathcal{N}(x_t + f(x_t, t; w), \sigma_t)$.*

إن مقياس **حجم البنى العميقة** (Deep Architectures) يعد عاملًا حاسمًا في أدائها، وكما رأينا في القسم § 3.7، فإن **النماذج اللغوية الضخمة** (LLMs) خصوصًا قد تتطلب **قدرات هائلة من الذاكرة والحوسبة** تتجاوز بكثير ما هو متاح في الأجهزة الشخصية العادية.

إن **تدريب مثل هذه النماذج من الصفر** يحتاج إلى موارد لا تتوافر عادة إلا لدى **الشركات الكبرى أو المؤسسات البحثية العامة**، ومع ذلك فقد تم تطوير تقنيات تمكّن من إجراء **الاستدلال (inference)** و**التخصيص لمهام معينة (adaptation)** حتى في ظل **قيود شديدة على الموارد**.

وتُعد القدرة على **تشغيل النماذج محليًا** بدلاً من الاعتماد على مزوّد سحابي أمرًا **مرغوبًا للغاية**، سواء **لأسباب مالية أو مرتبطة بالسرية والخصوصية**.

هندسة التعليمات (Prompt Engineering)

أبسط استراتيجية لتخصيص نموذج لغوي ضخم أو تحسين أدائه **ضمن ميزانية حوسبية محدودة** هي استخدام ما يُعرف بـ **هندسة التعليمات**، أي **صياغة بداية النص (prompt)** بعناية لتوجيه عملية التوليد الذاتية (autoregressive generation) نحو النتائج المرغوبة [Sahoo وآخرون، 2024].
في هذا الأسلوب، يتم **نقل جزء من المعلومات** التي كانت تُشفّر تقليديًا داخل **معاملات النموذج إلى مدخلاته النصية.**

وقد رأينا في القسم § 7.1 مثالًا بسيطًا على **التعلّم بالقليل من الأمثلة (few-shot prediction)**، حيث يمكن استخدام نموذج لغوي ضخم لتنفيذ مهمة **تصنيف نصي** دون الحاجة إلى إعادة تدريبه (fine-tuning). ومع ذلك، فإن **صياغة تعليمات طويلة ومعقدة** تتيح **تعميم هذه الاستراتيجية** لتشمل مهام أكثر تعقيدًا.

ولأن دور التعليمات هو **استغلال الانحيازات الإيجابية** التي اكتسبها النموذج من بيانات التدريب، فقد وُجد أن استراتيجيات غير متوقعة مثل **التصريح في النص بأن الإجابة مقدّمة من مختصّ محترف** يمكن أن **تحسّن النتائج بشكل ملحوظ** [Xu وآخرون، 2023].

ويُعد **طول السياق** (context length) — أي **عدد الرموز (الوحدات اللغوية، الكُليمات) (tokens)** التي يستطيع النموذج التعامل معها دفعةً واحدة — عاملًا أساسيًا في **كمية المعلومات التي يمكن تضمينها في التعليمات.**
إلا أن هذا الطول مقيد في الغالب بـ **الكلفة الحسابية العالية** للنماذج المعتمدة على **الاهتمام القياسي (standard attention)**، إذ **تنمو الكلفة تربيعياً مع طول السياق** (انظر § 4.8).

سلسلة التفكير (Chain of Thought)

نوع مدهش من أساليب هندسة التعليمات يهدف إلى جعل النموذج **ينتج خطواتٍ وسيطة للتفكير قبل تقديم الإجابة النهائية.**

تتكوّن هذه **السلسلة من التفكير** من **خطوات متتابعة**، تكون كل منها **أبسط من المسألة** الأصلية، ولذلك تكون **مماثلة بشكل أفضل لما تعلّمه النموذج** أثناء التدريب، كما أن تنبؤه بها يكون أكثر استقرارًا وحسمًا [Wei وآخرون، 2022؛ Kojima وآخرون، 2022].

(انظر الشكل 8.1 لمثال توضيحي.)

Q: Gina has 105 beans, she gives 23 beans to Bob, and prepares a soup with 53 beans. How many beans are left? A: There are 29 beans left.

Q: I prepare 53 pancakes, eat 5 of them and give 7 to Gina. I then prepare 26 more. How many pancakes are left? A: **27 pancakes are left.**

Q: Gina has 105 beans, she gives 23 beans to Bob, and prepares a soup with 53 beans. How many beans are left? A: Let's proceed step by step: Gina has 105 beans, she gives 23 beans to Bob (82 left), and prepares a soup with 53 beans (29 left). So there are 29 beans left.

Q: I prepare 53 pancakes, eat 5 of them and give 7 to Gina. I then prepare 26 more. How many pancakes are left? A: Let's proceed step by step: **53 pancakes, eat 5 of them (48 left), give 7 to Gina (41 left), prepare 26 more (67 left). So there are 67 pancakes left.**

Figure 8.1: *Example of a chain-of-thought to improve the response of the Llama-3-8B base model. In the two examples, the beginning of the text in normal font is the prompt, and the generated part is indicated in bold. The generation without chain-of-thought (top) leads to an incorrect answer, while the generation with it (bottom) generates a correct answer, by explicitly producing multiple simple arithmetic operations.*

التوليد المعزّز بالاسترجاع (Retrieval-Augmented Generation – RAG)

يمكن أيضًا توظيف **هندسة التعليمات (Prompt Engineering)** لربط نموذج لغوي بقاعدة معرفة خارجية. فهي تعمل هنا كـ **واجهة ذكية** تمكّن المستخدم من **طرح الأسئلة بلغة طبيعية** والحصول على **إجابات تجمع بين المعلومات المخزَّنة داخل النموذج والمعلومات المسترجعة من مصادر خارجية** [Lewis وآخرون، 2020].

في هذا النهج المسمّى **التوليد المعزّز بالاسترجاع (RAG)**، يُستخدم **نموذج تضمين (Embedding Model)** للبحث عن المستندات التي تكون **تضميناتها (Embeddings)** أكثر تقاربًا مع تضمين استعلام المستخدم. بعد ذلك، يُنشأ **نصّ تمهيدي (Prompt)** يجمع بين هذه المستندات المسترجعة وتعليماتٍ توضّح كيفية دمجها، ثم يتولى **النموذج التوليدي** إنتاج الإجابة النهائية للمستخدم بناءً على هذه المعلومات الموحّدة.

التكمية (Quantization)

رغم أن تدريب النماذج الضخمة أو تشغيلها على تدفقات متوازية متعددة **يتطلّب أجهزة حوسبة عالية الأداء**، فإن **تشغيل نموذج لغوي ضخم للاستخدام الفردي** غالباً ما يقتصر على **توليد تسلسل واحد في كل مرة (single-stream inference)**، مما يجعل **قيود الذاكرة وسرعة الوصول إليها** أهم بكثير من القدرة الحسابية بحد ذاتها.

كما ذُكر في § 2.1، فإن **المتغيّرات (parameters)** و**التنشيطات (activations)** و**التدرجات (gradients)** تُشفّر عادةً باستخدام **32 أو 16 بت**، وهي دقّة ضرورية أثناء التدريب لتمكين التغيّرات الصغيرة من التراكم التدريجي.
لكن أثناء **الاستدلال (inference)**، حيث تكون التنشيطات ناتجة عن مجموع عدد كبير من القيم، فإن تأثير التكمية يقلّ بفضل **تأثير المتوسط الإحصائي**. يزداد هذا الثبات كلما كان النموذج أضخم، بحيث يمكن **تكمية النماذج إلى 6 أو حتى 4 بت لكل معامل** دون فقد كبير في الأداء.

يُسهم هذا في **تقليل الذاكرة المطلوبة** بشكل كبير، كما **يحسّن سرعة الاستدلال** بوضوح. وقد أدى ذلك إلى تطوير أدوات برمجية تقوم بعملية **التكمية بعد التدريب (Post-Training Quantization)**، لتشغيل النماذج على أجهزة المستهلك العادية، مثل أداة **llama.cpp [Llama.cpp, 2023]**، التي تدعم **تنسيقات تكمية متعددة**، بحيث تُخصَّص مستويات دقّة مختلفة حسب نوع المصفوفة في النموذج اللغوي.
فعلى سبيل المثال، يمكن استخدام عدد أكبر من البتات لمعاملات **WV** في كتل الانتباه، ومعاملات الشبكات **الانتقالية (feed-forward blocks)**.

أحد أمثلة التكمية في **llama.cpp** هو تنسيق **Q4_1**.
يُقسّم هذا الأسلوب مصفوفة الأوزان إلى كتل فرعية من **32 عنصراً**، ولكل كتلة يتم تخزين **عامل مقياس d والانحياز m** بدقّة FP16، بينما يُكمَّم كل عنصر **x** إلى قيمة **q** من المجموعة {0,…,15} باستخدام 4 بت فقط، ليُعاد بعد ذلك فك التكمية بالتقريب $\tilde{x} = dq + m$.

كانت الكتلة الأصلية مشفَّرة في FP16 بحجم **64 بايت** (32 × 2 بايت)، بينما النسخة المكمَّاة تستهلك **20 بايت فقط** (4 بايت للمقياس والانحياز، و16 بايت للقيم المكمّاة).
ورغم هذا التقليص الكبير في الحجم، فإن الأداء لا يتأثر إلا **بشكل طفيف جدًا**، كما توضّحه النتائج في الشكل (8.2).

وهناك بديل آخر هو **التدريب الواعي بالتكمية (Quantization-Aware Training)**، حيث تُطبَّق التكمية أثناء **التمرير الأمامي (forward pass)**، لكن تبقى المعاملات والتدرجات مخزَّنة بدقّة عالية، وتُحتسب التدرجات في **التمرير الخلفي (backward pass)** كما لو لم تُطبَّق أي تكمية [Ma وآخرون، 2024].

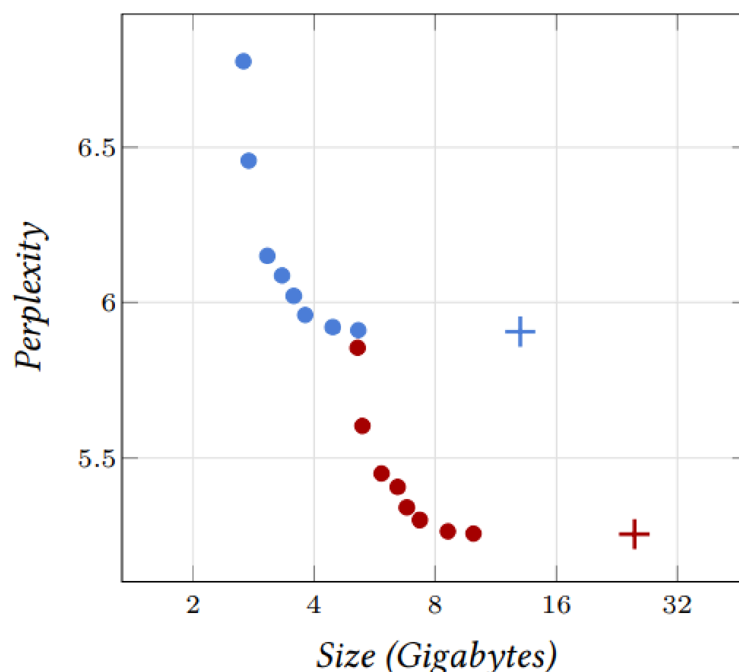


Figure 8.2: *Perplexity of quantized versions of the language models Llama-7B (blue) and 13B (red) [Touvron et al., 2023] on the wikitext corpus, as a function of the parameters' memory footprint. The crosses are the original FP16 models and the dots correspond to different levels of quantization with llama.cpp [Llama.cpp, 2023].*

المحوّلات الإضافية (المُكيّفات) (Adapters)

كما رأينا في § 3.6، تُعد **إعادة الضبط الدقيقة (Fine-tuning)** استراتيجية أساسية لإعادة استخدام النماذج المدربة مسبقًا. وبما أن الهدف منها هو إجراء **تعديلات طفيفة على نموذج قائم**، فقد طُوّرت تقنيات تضيف **مكوّنات جديدة صغيرة الحجم** تُعرف باسم **المحوّلات (Adapters)** إلى البنية الأصلية للنموذج، بينما **تُجمَّد جميع المعاملات الأصلية** ولا تُحدّث أثناء التدريب [Houlsby وآخرون، 2019].

الطريقة الأكثر شيوعًا اليوم هي **الضبط منخفض الرتبة (Low-Rank Adaptation – LoRA)**، وهي تضيف **تصحيحات منخفضة الرتبة** إلى بعض مصفوفات أوزان النموذج [Hu وآخرون، 2021].

من الناحية الرياضية، إذا كانت لدينا عملية خطية من الشكل:

$$X W^T$$

حيث X تمثل **مصفوفة التنشيطات (Activations)** ذات البُعد N×D (N عدد العينات في الدفعة، D بُعد السمات)، وW هي **مصفوفة الأوزان** ذات البُعد C×D،

فإن LoRA تستبدل هذه العملية بـ:

$$X (W + B A)^T$$

بحيث تكون A وB **مصفوفتين قابلتين للتدريب**، A ذات البُعد R×D وB ذات البُعد C×R، مع شرط أن **$R \ll \min(C, D)$**، أي أن الرتبة منخفضة مقارنة بأبعاد المصفوفة الأصلية.

يُزال W من مجموعة المعاملات القابلة للتدريب، وتُهيَّأ A بقيم عشوائية من التوزيع الغاوسي، بينما تُصفّر B، بحيث يبدأ الضبط الدقيق من نموذج يعطي **نفس مخرجات النموذج الأصلي تمامًا**.

عدد المعاملات الجديدة المطلوب ضبطها بهذه الطريقة لا يتجاوز عادة **عدة في المئة من عدد المعاملات الأصلية** للنموذج.

أما الإجراء القياسي لضبط **المحوّل (Transformer)** باستخدام LoRA فيقتصر على **مصفوفات الأوزان داخل كتل الانتباه (Attention Blocks)** فقط، بينما تُترك **الشبكات الانتقالية (Feed-Forward MLPs)** دون

تعديل. وقد استُخدمت الاستراتيجية نفسها بنجاح في **نماذج الانتشار (Diffusion Models)** لضبط كُتل الانتباه المسؤولة عن **التحكم النصي في توليد الصور.**

يسهم استخدام LoRA في **تقليل الذاكرة المطلوبة** بشكل كبير، لأن عدد المعاملات القابلة للتدريب أقل، وبالتالي يقلّ الحجم الذي يحتاجه **المحسِّن (Optimizer)** مثل Adam لتخزين المتوسطات الجارية لكل معامل. كما يُخفّض ذلك قليلًا من **تكلفة الحساب أثناء التمرير الخلفي (Backward Pass).**

وفي التطبيقات التجارية التي تتطلب **إصدارات كثيرة من النموذج المضبوط**، يمكن تخزين أزواج (A,B) الخاصة بكل إصدار **بشكل منفصل** عن النموذج الأصلي الذي يُخزَّن مرة واحدة فقط.

وفوق ذلك، يمكن **دمج التعديلات مباشرة داخل النموذج الأصلي** بمجرد حساب W ← W + A B، ليصبح النموذج الناتج مطابقًا تمامًا من حيث **البنية وعدد المعاملات أثناء الاستدلال (Inference)** للنموذج الأساسي.

وأخيرًا، رغم أن **التكمية (Quantization)** لا تؤثر عادة إلا قليلًا في دقة النموذج أثناء الاستدلال، فإن **الانحدار المتدرّج (Gradient Descent)** يتطلّب **دقة عالية** في تمثيل كل من التدرجات والمعاملات لتراكم التغيّرات الدقيقة.

ولهذا، يجمع أسلوب **QLoRA** بين **نموذج أساسي مكمّى** (لتقليل الذاكرة) **وتصحيحات LoRA غير مكمّاة** (للحفاظ على الدقة أثناء الضبط)، مما يوفّر توازنًا ممتازًا بين الأداء وكفاءة الذاكرة [Dettmers وآخرون، 2023].

بديلاً عن أساليب الضبط الدقيق والتوجيه المرئية في الأقسام السابقة، يتمثل في دمج نماذج متعددة ذات قدرات متنوعة في نموذج واحد، دون تدريب إضافي. يعتمد دمج النماذج على التوافق بين إصدارات متعددة مضبوطة بدقة من نموذج أساسي. أظهر إلهاركو وآخرون [2022] أن النماذج الناتجة عن الضبط الدقيق لنموذج CLIP الأساسي على عدة مجموعات بيانات لتصنيف الصور يمكن دمجها في فضاء المعاملات، حيث تُظهر خصائص الحساب المهامي (Task Arithmetic).

رسمياً، لنفرض أن θ هو متجه المعاملات لنموذج مُدرَّب مسبقاً، ولكل $t = 1,...,T$، لنفرض أن θt و $\tau t = \theta t - \theta$ هما المعاملات بعد الضبط الدقيق على المهمة t والبقايا المقابلة على التوالي. تُظهر التجارب أن النموذج ذا المعاملات $\theta + \tau 1 + \cdots + \tau T$ يُظهر قدرات متعددة المهام. وبالمثل، فإن طرح τt يُضعِف الأداء في المهمة المقابلة. تم تطوير أساليب لتقليل التداخل بين البقايا المختلفة وتحسين الأداء عندما يزداد عدد المهام [ياداف وآخرون، 2023؛ يو وآخرون، 2023].

بديل لدمج النماذج في فضاء المعاملات هو إعادة تركيب طبقاتها. يجمع أكيبا وآخرون [2024] بين دمج المعاملات وإعادة تركيب الطبقات، ويعتمدون على التحسين العشوائي للتعامل مع الانفجار التوافقي. تُظهر التجارب مع ثلاثة إصدارات مضبوطة بدقة من Mistral-7B [جيانغ وآخرون، 2023] أن الجمع بين استراتيجيتي الدمج هاتين يتفوق على كليهما.

الأجزاء المفقودة

من أجل الإيجاز، يتخطى هذا الكُتيب العديد من الموضوعات المهمة، وعلى وجه الخصوص:

الشبكات العصبية التكرارية

قبل أن تُظهر نماذج الانتباه أداءً أفضل، كانت الشبكات العصبية التكرارية (RNN) هي النهج المعياري للتعامل مع المتسلسلات الزمنية مثل النصوص أو عينات الصوت. تمتلك هذه المعماريات حالة داخلية مخفية يتم تحديثها في كل مرة تتم فيها معالجة مكون من المتسلسلة. مكوناتها الرئيسية هي طبقات مثل LSTM [هوشرايتر وشميدهوبر، 1997] أو GRU [تشو وآخرون، 2014].

يتمثل تدريب معمارية تكرارية في نشرها عبر الزمن، مما ينتج عنه تركيب طويل من المعاملات. وقد دفع هذا تاريخياً إلى تصميم تقنيات رئيسية تُستخدم الآن في المعماريات العميقة مثل المقوِّمات والبوابات، وهي شكل من أشكال الوصلات التخطّية التي يتم تعديلها ديناميكياً.

أحد العيوب الرئيسية للمعماريات التكرارية التقليدية هو أن بنية الحساب $x_{t+1} = f(x_t)$ تفرض معالجة متسلسلة الإدخال بشكل تسلسلي، مما يستغرق وقتاً يتناسب مع T. في المقابل، يمكن للمحولات (transformers)، على سبيل المثال، الاستفادة من الحساب المتوازي، مما ينتج عنه وقت ثابت إذا توفرت وحدات حوسبة كافية.

تتناول هذه المشكلة معماريات مثل QRNN [برادبري وآخرون، 2016]، وS4 [قو وآخرون، 2021]، أو Mamba [قو وداو، 2023]، التي تكون عملياتها التكرارية خطية بحيث يمكن حساب f^t نفسها، وبالتالي $x_t = f^t(x_0)$، بشكل متوازٍ، مما ينتج عنه وقت ثابت إذا لم تعتمد f على t و$\log T$ خلاف ذلك، مرة أخرى إذا توفرت وحدات حوسبة متوازية كافية.

المشفِّر الذاتي

المشفِّر الذاتي هو نموذج يربط إشارة إدخال، قد تكون عالية البُعد، بتمثيل كامن منخفض البُعد، ثم يربطه مرة أخرى بالإشارة الأصلية، مما يضمن الحفاظ على المعلومات. رأيناه في القسم 6.1 لإزالة التشويش، ولكن يمكن استخدامه أيضاً لاكتشاف معايرة منخفضة البُعد ذات مغزى تلقائياً لمنوع البيانات (data manifold).

المشفِّر الذاتي التبايني (VAE) الذي اقترحه كينجما وويلينج [2013] هو نموذج توليدي ببنية مماثلة. يفرض، من خلال دالة الخسارة، توزيعاً محدداً مسبقاً على التمثيل الكامن. وهذا يسمح، بعد التدريب، بتوليد عينات جديدة عن طريق أخذ عينات من التمثيل الكامن وفقاً لهذا التوزيع المفروض ثم الربط العكسي من خلال فك التشفير.

الشبكات التوليدية التنافسية

نهج آخر لنمذجة الكثافة هو الشبكات التوليدية التنافسية (GAN) التي قدّمها قودفيلو وآخرون [2014]. تجمع هذه الطريقة بين مولّد، يأخذ مدخلاً عشوائياً يتبع توزيعاً ثابتاً كمدخل وينتج إشارة منظمة مثل صورة، ومميّز، يأخذ عينة كمدخل ويتنبأ بما إذا كانت تأتي من مجموعة التدريب أو ما إذا كانت قد وُلِّدت بواسطة المولّد.

يُحسِّن التدريب المميّز لتقليل خسارة الإنتروبيا المتقاطعة المعيارية، والمولّد لتعظيم خسارة المميّز. يمكن إثبات أنه عند التوازن، ينتج المولّد عينات لا يمكن تمييزها عن البيانات الحقيقية. في الممارسة العملية، عندما يتدفق التدرج عبر المميّز إلى المولّد، فإنه يُبلِغ الأخير عن الإشارات التي يستخدمها المميّز والتي تحتاج إلى معالجة.

الشبكات العصبية البيانية

تتطلب العديد من التطبيقات معالجة إشارات غير منظمة بانتظام على شبكة. على سبيل المثال، البروتينات، والشبكات ثلاثية الأبعاد، والمواقع الجغرافية، أو التفاعلات الاجتماعية منظمة بشكل أكثر طبيعية كرسوم بيانية. الشبكات الالتفافية المعيارية أو حتى نماذج الانتباه غير ملائمة بشكل جيد لمعالجة مثل هذه البيانات، والأداة المفضلة لمثل هذه المهمة هي الشبكات العصبية البيانية (GNN) [سكارسيلي وآخرون، 2009].

تتكون هذه النماذج من طبقات تحسب التنشيطات عند كل رأس من خلال الجمع الخطي للتنشيطات الموجودة في الرؤوس المجاورة المباشرة له. هذه العملية مشابهة جداً للالتفاف المعياري، باستثناء أن بنية البيانات لا تعكس أي معلومات هندسية مرتبطة بمتجهات السمات التي تحملها.

التدريب ذاتي الإشراف

كما ذُكر في القسم 7.1، على الرغم من تدريبها فقط للتنبؤ بالكلمة التالية، فإن نماذج اللغة الكبيرة المدرَّبة على مجموعات بيانات كبيرة غير مُعنونة مثل GPT (انظر القسم 5.3) قادرة على حل مهام متنوعة، مثل تحديد الدور النحوي لكلمة، أو الإجابة عن الأسئلة، أو حتى الترجمة من لغة إلى أخرى [رادفورد وآخرون، 2019].

تشكل مثل هذه النماذج فئة واحدة من فئة أكبر من الأساليب التي تندرج تحت اسم التعلم ذاتي الإشراف، وتحاول الاستفادة من مجموعات البيانات غير المُعنونة [باليستريرو وآخرون، 2023]. المبدأ الأساسي لهذه الأساليب هو تعريف مهمة لا تتطلب تسميات ولكنها تستلزم تمثيلات سمات مفيدة للمهمة الحقيقية محل الاهتمام، والتي توجد لها مجموعة بيانات مُعنونة صغيرة. في الرؤية الحاسوبية، على سبيل المثال، يمكن تحسين سمات الصورة بحيث تكون ثابتة (invariant) لتحويلات البيانات التي لا تغير المحتوى الدلالي للصورة، بينما تكون غير مترابطة إحصائياً [زبونتار وآخرون، 2021].

في كل من معالجة اللغة الطبيعية والرؤية الحاسوبية، تمثل الاستراتيجية العامة القوية في تدريب نموذج لاستعادة أجزاء من الإشارة التي تم إخفاؤها [ديفلين وآخرون، 2018؛ تشو وآخرون، 2021].

جدول المصطلحات الإنجليزية والعربية

المقابل العربي	**المصطلح الإنجليزي**
الالتفاف أحادي البُعد	1D convolution
الالتفاف ثنائي البُعد	2D convolution
التنشيط/ التفعيل	activation
دالة التنشيط	activation function
خريطة التنشيط	activation map
المُكيِّف	adapter
عملية خطية	affine operation

artificial neural network	الشبكة العصبية الاصطناعية (الشعصبونة)
attention operator	معامل الانتباه
autoencoder	المشفِّر الذاتي
denoising autoencoder	المشفِّر الذاتي لإزالة التشويش
Autograd	الاشتقاق التلقائي
autoregressive model	النموذج الانحداري الذاتي (النكوصي)
average pooling	التجميع بالمتوسط
backpropagation	الانتشار العكسي
backward pass	المرور العكسي (العودة)

basis function regression	انحدار (انكفاء) دالة الأساس
batch	دفعة
batch normalization	التنظيم الدفعي (المسوِّي)
Bellman equation	معادلة بيلمان
bias vector	متجه الانحياز
BPE Byte Pair Encoding	ترميز زوج البايت
cache memory	الذاكرة المخبئية
capacity	السعة
causal model	النموذج السببي

(chain rule (derivative	قاعدة السلسلة (الاشتقاق)
(chain rule (probability	قاعدة السلسلة (الاحتمال)
chain-of-thought	سلسلة التفكير
channel	قناة
checkpointing	إنشاء نقاط التفتيش (محفوظة)
classification	التصنيف
CLIP (Contrastive Language-Image (Pre-training	التدريب المسبق التبايُنيّ للغة والصورة
CLS token	رمز التصنيف (المُمَثِّل)
computational cost	التكلفة الحسابية

context size	حجم السياق
contrastive loss	خسارة تباينية
convolution	الالتفاف
convolutional layer	الطبقة الالتفافية
convolutional network	الشبكة الالتفافية
cross-attention block	كتلة الانتباه التقاطعي
cross-entropy	الإنتروبيا المتقاطعة (الاعتلاج، الأنطروب)
data augmentation	تعزيز البيانات (تذييل)
deep learning	التعلم العميق

(Deep Q-Network (DQN	شبكة Q العميقة
density modeling	نمذجة الكثافة
depth	العمق
diffusion model	نموذج الانتشار (التغبيش، التشويش)
dilation	التوسيع
discriminator	المميِّز/ المفرّق
downscaling residual block	كتلة البقايا للتصغير
downstream task	المهمة اللاحقة
dropout	الإسقاط

embedding layer	طبقة التضمين
epoch	حقبة/لفّة
equivariance	التباين المشترك
feed-forward block	لبنة التغذية الأمامية
few-shot prediction	التنبؤ بعينات قليلة
filter	مرشّح
fine-tune	الضبط الدقيق
fine-tuning	الضبط الدقيق
flops	عمليات الفاصلة العائمة

forward pass	المرور الأمامي
foundation model	النموذج الأساسي
FP32	الفاصلة العائمة 32 بت
framework	إطار العمل
(GAN (Generative Adversarial Networks	الشبكات التوليدية التنافسية
GELU	دالة التنشيط GELU
generator	المولِّد
(GNN (Graph Neural Network	الشبكة العصبية البيانية
(GPT (Generative Pre-trained Transformer	المحِوّل التوليدي المُدَرَّب مسبقاً

(GPU (Graphical Processing Unit	وحدة معالجة الرسومات
gradient descent	الانحدار التدريجي
gradient norm clipping	قص معيار التدرج
gradient step	خطوة التدرج
ground truth	الحقيقة الأرضية
hidden layer	الطبقة المخفية
hidden state	الحالة المخفية
hyper parameter	المعامل الفائق
(hyperbolic tangent (Tanh	الظل الزائدي

image processing	معالجة الصور
image synthesis	توليد الصور
inductive bias	الانحياز الاستقرائي
invariance	الثبات / عدم التباين
kernel size	حجم النواة
key	المفتاح
(Large Language Model (LLM	نموذج اللغة الكبير
layer	طبقة
attention layer	طبقة الانتباه

fully connected layer	الطبقة كاملة الاتصال
linear layer	الطبقة الخطية
layer normalization	تطبيع الطبقة
Leaky ReLU	وحدة التقويم الخطي المسرِّبة
learning rate	معدل التعلم
learning rate schedule	جدول معدل التعلم
LeNet	شبكة LeNet
local minimum	الحد الأدنى المحلي
logit	اللوجيت

(LoRA (Low-Rank Adaptation	التكيف منخفض الرتبة
loss	الخسارة
machine learning	التعلم الآلي
(Markovian Decision Process (MDP	عملية قرار ماركوف
Markovian property	خاصية ماركوف
max pooling	التجميع بالقيمة العظمى
mean squared error	متوسط مربع الخطأ
memory requirement	متطلبات الذاكرة
memory speed	سرعة الذاكرة

metric learning	تعلم المقاييس
(MLP (multi-layer perceptron	الإدراك متعدد الطبقات
model	نموذج
model merging	دمج النماذج
(Natural Language Processing (NLP	معالجة اللغة الطبيعية
non-linearity	اللاخطية
normalizing layer	الطبقة المطّبِعة
object detection	كشف الأجسام
overfitting	الإفراط في التلاؤم

padding	الحشو
parameter	معامل
parametric model	النموذج الموزون
peak performance	الأداء الذروة
perplexity	الحيرة
policy	السياسة
optimal policy	السياسة المثلى
pooling	التجميع
positional encoding	الترميز الموضعي

Post-Training Quantization	الكمّيّة بعد التدريب
posterior probability	الاحتمال البعدي
pre-trained model	النموذج المُدرَّب مسبقاً
prompt	التوجيه
prompt engineering	هندسة التوجيه
quantization	الكمّيّة
Quantization-Aware Training	التدريب الواعي بالكمّيّة
query	الاستعلام
(RAG (Retrieval-Augmented Generation	التوليد المعزز بالاسترجاع

random initialization	التهيئة العشوائية
receptive field	الحقل الاستقبالي
(rectified linear unit (ReLU	وحدة التقويم الخطي
(recurrent neural network (RNN	الشبكة العصبية التكرارية
regression	الانحدار
(Reinforcement Learning (RL	التعلم المعزز
RLHF (Reinforcement Learning from Human (Feedback	التعلم المعزز من التغذية الراجعة البشرية
residual block	كتلة البقايا
residual connection	وصلة البقايا

residual network	الشبكة البقائية
ResNet-50	شبكة ResNet-50
return	العائد
reversible layer	الطبقة العكوسة
scaling laws	قوانين التوسع
self-attention block	كتلة الانتباه الذاتي
self-supervised learning	التعلم ذاتي الإشراف
semantic segmentation	التجزئة الدلالية
(SGD (stochastic gradient descent	الانحدار التدريجي العشوائي

(Single Shot Detector (SSD	الكاشف ذو اللقطة الواحدة
skip connection	الوصلة التخطّية
softargmax	دالة softargmax
softmax	دالة softmax
speech recognition	التعرف على الكلام
stride	الخطوة
supervised learning	التعلم المُشرَف
Task Arithmetic	الحساب المهامي
tensor	موتِّر

tensor cores	أنوية الموتِّرات
(Tensor Processing Unit (TPU	وحدة معالجة الموتِّرات
test set	مجموعة الاختبار
text synthesis	توليد النص
token	رمز
tokenizer	المُرمِّز
trainable parameter	المعامل القابل للتدريب
training	التدريب
training set	مجموعة التدريب

Transformer	المحوِّل
transposed convolution	الالتفاف المنقول
underfitting	قصور التلاؤم
universal approximation theorem	مبرهنة التقريب الشامل
unsupervised learning	التعلم غير المُشرَف
(VAE (variational autoencoder	المشفِّر الذاتي التبايني
validation set	مجموعة التحقق
value	القيمة
vanishing gradient	التدرج المتلاشي

variational bound	الحد التبايني ّ
(Vision Transformer (ViT	محوِّل الرؤية
vocabulary	المفردات
weight	الوزن
weight decay	اضمحلال الوزن
weight matrix	مصفوفة الأوزان
zero-shot prediction	التنبؤ بدون عينات